Synthese Library 376

Vincent C. Müller *Editor*

Fundamental Issues of Artificial Intelligence

Springer

Synthese Library

Studies in Epistemology, Logic, Methodology, and Philosophy of Science

Volume 376

More information about this series at http://www.springer.com/series/6607

Vincent C. Müller
Editor

Fundamental Issues of Artificial Intelligence

Springer

Editor
Vincent C. Müller
Future of Humanity Institute
Department of Philosophy
& Oxford Martin School
University of Oxford
Oxford, UK

Anatolia College/ACT
Thessaloniki, Greece
http://orcid.org/0000-0002-4144-4957
http://www.sophia.de

Synthese Library
ISBN 978-3-319-26483-7 ISBN 978-3-319-26485-1 (eBook)
DOI 10.1007/978-3-319-26485-1

Library of Congress Control Number: 2016930294

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Editorial Note

The papers in this volume result from the second conference on the *Philosophy and Theory of Artificial Intelligence* (PT-AI) 21-22.09.2013 that I organised in Oxford where I was a research fellow – for details on the conference, see http://www.pt-ai.org/.

For this conference, we had 103 extended abstract submissions by the deadline, which were reviewed double-blind by at least two referees. Thirty-four submissions, i.e. 33 %, were accepted for presentation. In a second reviewing phase, submitted full papers plus papers from invited speakers at the conference and papers from additional invited authors were openly reviewed and discussed between all these authors. The second reviewing phase resulted in 9 further rejections, so we now have 27 submitted papers, 3 from invited speakers and 3 invited ones, for a total of 33. Finally, the volume was reviewed by the publisher, which resulted in further revisions. We are grateful for all the hard work that went into this volume. Unfortunately, this process of reviewing, inviting additional authors, revising, re-reviewing, etc., took much longer than anticipated so we submit the final version to the publisher more than one and a half years after the conference.

Anatolia College/ACT, Thessaloniki, Greece Vincent C. Müller
19 June 2015

Contents

Chapter 1
New Developments in the Philosophy of AI

Vincent C. Müller

Abstract The philosophy of AI has seen some changes, in particular: (1) AI moves away from cognitive science, and (2) the long term risks of AI now appear to be a worthy concern. In this context, the classical central concerns – such as the relation of cognition and computation, embodiment, intelligence & rationality, and information – will regain urgency.

Keywords AI risk • Cognitive science • Computation • Embodiment • Philosophy of AI • Philosophy of artificial intelligence • Rationality • Superintelligence

1.1 Getting Interesting Again?

We set the framework for this conference broadly by these questions: "What are the necessary conditions for artificial intelligence (if any); what are sufficient ones? What do these questions relate to the conditions for intelligence in humans and other natural agents? What are the ethical and societal problems that artificial intelligence raises, or will raise?" – thus far, this was fairly similar to the themes for the 2011 conference (Müller 2012, 2013).

This introduction is also a meditation on a remark by one of our keynote speakers, Daniel Dennett, who wrote on Twitter: "In Oxford for the AI conference. I plan to catch up on the latest developments. It's getting interesting again." (@danieldennett 19.09.2013, 11:05 pm). If Dennett thinks "it's getting interesting" that is good news, and it is significant that he remarks that this interest appears *again*.

In the following year, the AAAI invited me to speak about "What's Hot in the Philosophy of AI?" (their title) – so the organization of AI researchers around the world also thinks it might be worthwhile to have a look at philosophy *again*. And

V.C. Müller (✉)
Future of Humanity Institute, Department of Philosophy & Oxford Martin School, University of Oxford, Oxford, UK

Anatolia College/ACT, Thessaloniki, Greece
e-mail: vmueller@act.edu; http://www.sophia.de

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, Synthese Library 376, DOI 10.1007/978-3-319-26485-1_1

indeed, one of the major topics in the AAAI plenary discussion was the social impact of AI; the president of AAAI has now made ethics an 'official' topic of concern (Ditterich and Horowitz 2015).

So, there are indications that 'philosophy of artificial intelligence' might have an impact, again. I think there are two major changes that make this the case: The changing relation to cognitive science and the increasing urgency of ethical concerns.

1.2 AI & CogSci: A Difficult Marriage

The traditional view of AI and Cognitive Science has been that they are two sides of the same coin, two efforts that require each other or even the same effort with two different methods. The typical view in the area of 'good old fashioned AI' (GOFAI, as Haugeland called it) until the 1980s was that the empirical discipline of cognitive science finds how natural cognitive systems (particularly humans) work, while the engineering discipline of AI tests the hypotheses of cognitive science and uses them for progress in its production of artificial cognitive systems. This marriage was thus made on the basis of a philosophical analysis of joint assumptions – so philosophy served as the 'best man'.

This collaboration was made possible, or we at least facilitated, by the by classical 'machine functionalism, going back to (Putnam 1960) and nicely characterized by Churchland: "What unites them [the cognitive creatures] is that [. . .] they are all computing the same, or some part of the same abstract <<sensory input, prior state>, <motor output, subsequent state>>*function*." (Churchland 2005: 333). If cognition is thus a computational process over symbolic representation (this thesis is often called 'computationalism') then computation can be discovered by cognitive science and then implemented by AI in an artificial computational system. This was typically complemented by a view of cognition as central 'control' of an agent that follows a structure of sense-model-plan-act; that rationally 'selects' an action, typically given some utility function. – All these components have been the target of powerful criticism over the years.

1.3 After GOFAI, "What's Hot in the Philosophy of AI?"

Two factors are different now from the way things looked only 10 or 20 years ago: (a) We now have much more impressive technology, and (b) we have a different cognitive science. The result is, or so I will argue, that we get a new theory of AI and new ethics of AI.

It currently looks like after the cold 'AI winter' in the 1980s and 90s we are already through a spring and staring a nice and warm summer with AI entering

the mainstream of computing and AI products – even if much of this success does not carry the name of 'artificial intelligence' any more. This is a version of the well-known 'AI curse': in the formulation known as 'Larry Tesler's Theorem' (ca. 1970): "Intelligence is whatever machines haven't done yet." What is successful in AI takes a different name (e.g. 'machine learning'), and what is left for the old name are the currently impossible problems and the long–term visions.

A lot of classical AI problems are now solved and even thought trivial (e.g. real-life character recognition); robotics is now moving beyond the classic DDD problems (dirty, dangerous, dull). It appears that this is largely due to massively improved computing resources (processing speed and the ability to handle large data sets), the continued 'grind' forward towards better algorithms and a certain focus on feasible narrower problems. It does not seem to be due to massive new deep insight.

What does this mean for our marriage? Is there a divorce in the offing? The cognitive science side has largely learned to live with in separation – not quite a divorce but a more independent life. There is no more the assumption that cognition must be algorithmic (computational) symbol processing but rather a preference for broadly computational models. A strong emphasis on empirical work supports a tendency of cognitive science to undergo a metamorphosis from a multidisciplinary enterprise to another word for cognitive psychology. Cognitive science now involves embodied theories, dynamic theories, etc. – and it tends to find its own path now, not as adjunct to AI.

1.4 Ethics (Big and Small)

It has always been clear that AI, esp. higher level AI, will have an 'impact on society' (e.g. surveillance, jobs, weapons & war, care, ...) and that some of that impact is undesirable, perhaps requiring policy interference. There is also the impact on the self-image of humans that makes AI, and especially robotics, have such a powerful impression on people who care a lot less about other new technologies. This is what I would call 'small ethics', the kind of ethics that concerns impact on a relatively small scale.

There is also 'big ethics' of AI that asks about a very large impact on society, and on the human kind. A discussion of this issue is relatively new in academic circles. Stuart Russell, one of our keynote speakers, had called it the question "what if we succeed?" (at IJCAI 2013) – (Bostrom 2014; Russell et al. 2015).

If the results of the paper by Bostrom and myself in this volume are to be believed (Müller and Bostrom 2016), then experts estimate the probability of achieving 'high level machine intelligence' to go over 50 % by 2040–2050, over 90 % by 2075. Broadly, this will happen soon enough to think about it now, especially since 30 % of the same experts think, that the outcome of achieving such machine intelligence will be 'bad' or 'very bad' for humanity.

I expect that this theme will create much discussion and interest, and that its speculation about what can be and what will be forces a return to the 'classical' themes of the philosophy of AI, including the relation of AI and cognitive science.

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Churchland, P. M. (2005). Functionalism at forty: A critical retrospective. *Journal of Philosophy, 102*(1), 33–50.

Ditterich, T., & Horowitz, E. Benefits and risks of artificial intelligence, *medium.com*. https://medium.com/@tdietterich/benefits-and-risks-of-artificial-intelligence-460d288cccf3%3E. Accessed 23 Jan 2015.

Müller, V. C. (ed.). (2012) *Theory and philosophy of AI* (Minds and machines, 22/2- Special volume). Springer.

Müller, V. C. (ed.). (2013). *Theory and philosophy of artificial intelligence* (SAPERE, vol. 5). Berlin: Springer.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 553–570). Synthese Library; Berlin: Springer.

Putnam, H. (1960). Minds and machines. In *Mind, language and reality. Philosophical Papers, Volume II* (pp. 362–385). Cambridge: Cambridge University Press 1975).

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. http://futureoflife.org/static/data/documents/research_priorities.pdf%3E

Part I
Computing

Chapter 2
Rationality and Intelligence: A Brief Update

Stuart Russell

Abstract The long-term goal of AI is the creation and understanding of intelligence. This requires a notion of intelligence that is precise enough to allow the cumulative development of robust systems and general results. The concept of *rational agency* has long been considered a leading candidate to fulfill this role. This paper, which updates a much earlier version (Russell, Artif Intell 94:57–77, 1997), reviews the sequence of conceptual shifts leading to a different candidate, *bounded optimality*, that is closer to our informal conception of intelligence and reduces the gap between theory and practice. Some promising recent developments are also described.

Keywords Rationality • Intelligence • Bounded rationality • Metareasoning

2.1 Artificial Intelligence

AI is a field whose ultimate goal has often been somewhat ill-defined and subject to dispute. Some researchers aim to emulate human cognition, others aim at the creation of intelligence without concern for human characteristics, and still others aim to create useful artifacts without concern for abstract notions of intelligence.

My own motivation for studying AI is to create and understand intelligence as a general property of systems, rather than as a specific attribute of humans. I believe this to be an appropriate goal for the field as a whole, and it certainly includes the creation of useful artifacts—both as a spin-off from and a driving force for technological development. The difficulty with this "creation of intelligence" view, however, is that it presupposes that we have some productive notion of what intelligence is. Cognitive scientists can say "Look, my model correctly predicted

S. Russell (✉)
Computer Science Division, University of California, Berkeley, CA 94720, USA
e-mail: russell@cs.berkeley.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_2

this experimental observation of human cognition," and artifact developers can say "Look, my system is worth billions of euros," but few of us are happy with papers saying "Look, my system is intelligent."

A definition of intelligence needs to be *formal*—a property of the system's input, structure, and output—so that it can support analysis and synthesis. The Turing test does not meet this requirement, because it references an informal (and parochial) human standard. A definition also needs to be *general*, rather than a list of specialized faculties—planning, learning, game-playing, and so on—with a definition for each. Defining each faculty separately presupposes that the faculty is *necessary* for intelligence; moreover, the definitions are typically not composable into a general definition for intelligence.

The notion of *rationality* as a property of *agents*—entities that perceive and act—is a plausible candidate that may provide a suitable formal definition of intelligence. Section 2.2 provides background on the concept of agents. The subsequent sections, following the development in Russell (1997), examine a sequence of definitions of rationality from the history of AI and related disciplines, considering each as a predicate P that might be applied to characterize systems that are intelligent:

- P_1: *Perfect rationality*, or the capacity to generate maximally successful behaviour given the available information.
- P_2: *Calculative rationality*, or the in-principle capacity to compute the perfectly rational decision given the initially available information.
- P_3: *Metalevel rationality*, or the capacity to select the optimal combination of computation-sequence-plus-action, under the constraint that the action must be selected by the computation.
- P_4: *Bounded optimality*, or the capacity to generate maximally successful behaviour given the available information and computational resources.

For each P, I shall consider three simple questions. First, are P-systems interesting, in the sense that their behaviour is plausibly describable as intelligent? Second, could P-systems ever exist? Third, to what kind of research and technological development does the study of P-systems lead?

Of the four candidates, P_4, bounded optimality, comes closest to meeting the needs of AI research. It is more suitable than P_1 through P_3 because it is a real problem with real and desirable solutions, and also because it satisfies some essential intuitions about the nature of intelligence. Some important questions about intelligence can only be formulated and answered within the framework of bounded optimality or some relative thereof.

2.2 Agents

In the early decades of AI's history, researchers tended to define intelligence with respect to specific tasks and the internal processes those tasks were thought to require in humans. Intelligence was believed to involve (among other things) the

ability to understand language, the ability to reason logically, and the ability to solve problems and construct plans to satisfy goals. At the core of such capabilities was a store of knowledge. The standard conception of an AI system was as a sort of *consultant*: something that could be fed information and could then answer questions. The output of answers was not thought of as an *action* about which the AI system had a choice, any more than a calculator has a choice about what numbers to display on its screen given the sequence of keys pressed.

The view that AI is about building intelligent *agents*—entities that sense their environment and act upon it—became the mainstream approach of the field only in the 1990s (Russell and Norvig 1995; Dean et al. 1995), having previously been the province of specialized workshops on "situatedness" and "embeddedness". The "consultant" view is a special case in which answering questions is a form of acting—a change of viewpoint that occurred much earlier in the philosophy of language with the development of speech act theory. Now, instead of simply giving answers, a consulting agent could refuse to do so on the grounds of privacy or promise to do so in return for some consideration. The agent view also naturally encompasses the full variety of tasks and platforms—from robots and factories to game-playing systems and financial trading systems—in a single theoretical framework.

What matters about an agent is what it *does*, not how it does it. An agent can be defined mathematically by an *agent function* that specifies how an agent behaves under all circumstances. More specifically, let $\mathbf{O}$ be the set of percepts that the agent can observe at any instant (with $\mathbf{O}^*$ being the set of observation sequences of any length) and $\mathbf{A}$ be the set of possible actions the agent can carry out in the external world (including the action of doing nothing). The agent function is a mapping $f : \mathbf{O}^* \rightarrow \mathbf{A}$. This definition is depicted in the upper half of Fig. 2.1.

As we will see in Sect. 2.3, rationality provides a normative prescription for agent functions and does not specify—although it does constrain—the process by which the actions are selected. Rather than *assume* that a rational agent must, for example, reason logically or calculate expected utilities, the arguments for (Nilsson 1991) or against (Agre and Chapman 1987; Brooks 1989) the inclusion of such cognitive

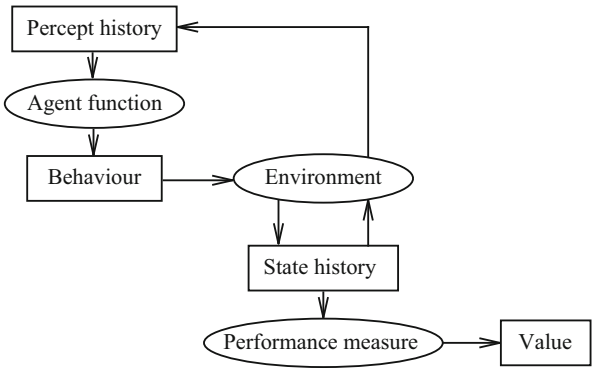


Fig. 2.1 The agent receives percepts from the environment and generates a behaviour which in turn causes the environment to generate a state history. The performance measure evaluates the state history to arrive at the value of the agent

faculties must justify their position on the grounds of efficacy in representing a desirable agent function. A designer of agents has, a priori, complete freedom in choosing the specifications, boundaries, and interconnections of subsystems, as long they they compose to form a complete agent. In this way one is more likely to avoid the "hallucination" problem that arises when the fragility of a subsystem is masked by having an intelligent human providing input to it and interpreting its outputs.

Another important benefit of the agent view of AI is that it connects the field directly to others that have traditionally looked on the embedded agent as a natural topic of study, including economics, operations research, control theory, and even evolutionary biology. These connections have facilitated the importation of technical ideas (Nash equilibria, Markov decision processes, and so on) into AI, where they have taken root and flourished.

2.3 Perfect Rationality

So which agent functions are intelligent? Clearly, doing the right thing is more intelligent that doing the wrong thing. The rightness of actions is captured by the notion of rationality: informally, an action is rational to the extent that is consistent with the agent's goals (or the task for which it was designed), from the point of view of the information possessed by the agent.

Rationality is, therefore, always understood relative to the agent's ultimate goals. These are expressed mathematically by a performance measure U on sequences of environment states. Let $V(f, \mathbf{E}, U)$ denote the expected value according to U obtained by an agent function f in environment class $\mathbf{E}$, where (for now) we will assume a probability distribution over elements of $\mathbf{E}$. Then a perfectly rational agent is defined by an agent function f_{opt} such that

$$f_{\text{opt}} = \text{argmax}_f V(f, \mathbf{E}, U) \tag{2.1}$$

This is just a fancy way of saying that the best agent does the best it can. The point is that perfectly rational behaviour is a well-defined function of the *task environment* fixed by $\mathbf{E}$ and U.

Turning to the three questions listed in Sect. 2.1: Are perfectly rational agents interesting things to have? Yes, certainly—if you have one handy, you prefer it to any other agent. A perfectly rational agent is, in a sense, perfectly intelligent. Do they exist? Alas no, except for very simple task environments, such as those in which *every* behavior is optimal (Simon 1958). Physical mechanisms take time to perform computations, while real-world decisions generally correspond to intractable problem classes; imperfection is inevitable.

Despite their lack of existence, perfectly rational agents have, like imaginary numbers, engendered a great deal of interesting research. For example, economists prove nice results about economies populated by them and game-theoretic mechanism designers much prefer to assume perfect rationality on the part of each agent.

Far more important for AI, however, was the reduction from a global optimization problem (Eq. 2.1) to a local one: from the perfect rationality of agents to the perfect rationality of individual actions. That is, a perfectly rational agent is one that repeatedly picks an action that maximizes the expected utility of the next state. This reduction involved three separate and largely unconnected results: the axiomatic utility theory of von Neumann and Morgenstern (1944) (which actually takes for granted the agent's ability to express preferences between distributions over immediate outcomes), Bellman's 1957 theory of sequential decisions, and Koopmans' 1972 analysis of preferences over time in the framework of multiattribute utility theory (Keeney and Raiffa 1976).

While utility is central to the decision-theoretic notion of perfect rationality, *goals* are usually considered to define the task for a logic-based agent: according to Newell (1982), such an agent is perfectly rational if each action is part of a plan that will achieve one of the agent's goals. There have been attempts to define goals in terms of utilities, beginning with Wellman and Doyle (1991), but difficulties remain because goals are essentially incomplete as task specifications. They do not specify what to do when goal achievement cannot be guaranteed, or when goals conflict, or when several plans are available for achieving a goal, or when the agent has achieved all its goals. It may be better to interpret goals not as primary definitions of the agent's task but as subsidiary devices for focusing computational effort with an overall decision-theoretic context. For example, someone moving to a new city may, after weighing many alternatives and tradeoffs under uncertainty, settle on the goal of buying a particular apartment and thereafter focus their deliberations on finding a plan to achieve that goal, to the exclusion of other possibilities. At the moment we do not have a good understanding of goal formation by a decision-theoretic agent, but it is clear that such behavior cannot be analyzed within the framework of perfect rationality.

As discussed so far, the framework does not say where the beliefs and the performance measure reside—they could be in the head of the designer or of the agent itself. If they are in the designer's head, the designer has to do all the work to build the agent function, anticipating all possible percept sequences. If they are in the agent's head, the designer can delegate the work to the agent; for example, in the setting of reinforcement learning, it is common to equip the agent with a fixed capacity to extract a distinguished reward signal from the environment, leaving the agent to learn the corresponding utility function on states. The designer may also equip the agent with a prior over environments (Carnap 1950), leaving the agent to perform Bayesian updating as it observes the particular environment it inhabits. Solomonoff (1964) and Kolmogorov (1965) explored the question of universal priors over computable environments; universality, unfortunately, leads to undecidability of the learning problem. Hutter (2005) makes an ambitious attempt to define a universal yet computable version of perfect rationality, but does not pretend to provide the instantaneous decisions required for an actual P_1-system; instead, this work belongs in the realm of P_2-systems, or calculatively rational agents.

Perhaps the biggest open question for the theory of perfect rationality lies in its extension from single-agent to multi-agent environments. Game theorists

have proposed many *solution concepts*—essentially, definitions of admissible strategies—but have not identified one that yields a unique recommendation (up to tie-breaking) for what to do (Shoham and Leyton-Brown 2009).

2.4 Calculative Rationality

The theory of P_1, perfect rationality, says nothing about implementation; P_2, calculative rationality, on the other hand, is concerned with programs for computing the choices that perfect rationality stipulates.

To discuss calculative rationality, then, we need to discuss programs. The agent's decision-making system can be divided into the *machine* M, which is considered fixed, and the *agent program* l, which the designer chooses from the space $\mathscr{L}_M$ of all programs that the machine supports. (M need not be a raw physical computer, of course; it can be a software "virtual machine" at any level of abstraction.) Together, the machine M and the agent program l define an agent function $f = Agent(l, M)$, which, as noted above, is subject to evaluation. Conversely, l is an *implementation* of the agent function f on M; there may, of course, be many such implementations, but also, crucially, there may be none (see Sect. 2.6).

It is important to understand the distinction between an agent program and the agent function it implements. An agent program may receive as input the current percept, but also has internal state that reflects, in some form, the previous percepts. It outputs actions when they have been selected. From the outside, the behaviour of the agent consists of the selected actions *interspersed with inaction* (or whatever default actions the machine generates). Depending on how long the action selection takes, many percepts may go by unnoticed by the program.

Calculative rationality is displayed by programs that, *if executed infinitely fast*, would result in perfectly rational behaviour. That is, at time t, assuming it is not already busy computing its choice for some previous time step, the program computes the value $f_{\text{opt}}([o_1, \ldots, o_t])$.

Whereas perfect rationality is highly desirable but does not exist, calculative rationality often exists—its requirements can be fulfilled by real programs for many settings—but it is not necessarily a desirable property. For example, a calculatively rational chess program will choose the "right" move, but may take 10^{50} times too long to do so.

The pursuit of calculative rationality has nonetheless been the main activity of theoretically well-founded research in AI; the field has been filling in a table whose dimensions are the various environment properties (deterministic or stochastic, fully or partially observable, discrete or continuous, dynamic or static, single-agent or multi-agent, known or unknown) for various classes of representational formalisms (atomic, propositional, or relational). In the logical tradition, planning systems and situation-calculus theorem-provers satisfy the conditions of calculative rationality for discrete, fully observable environments; moreover, the power of first-order logic renders the required knowledge practically expressible for a wide range of problems.

In the decision-theoretic tradition, there are calculatively rational agents based on algorithms for solving fully or partially observable Markov decision processes, defined initially atomic by atomic formalisms (e.g., transition matrices), later by propositional representations (e.g., dynamic Bayesian networks), and now by first-order probabilistic languages (Srivastava et al. 2014). For continuous domains, stochastic optimal control theory (Kumar and Varaiya 1986) has solved some restricted classes of problems, while many others remain open.

In practice, neither the logical nor the decision-theoretic traditions can avoid the intractability of the decision problems posed by the requirement of calculative rationality. One response, championed by Levesque (1986), is to rule out sources of exponential complexity in the representations and reasoning tasks addressed, so that calculative and perfect rationality coincide—at least, if we ignore the little matter of polynomial-time computation. The accompanying research results on tractable sublanguages are perhaps best seen as indications of where complexity may be an issue rather than as a solution to the problem of complexity, since real-world problems usually require exponentially large representations under the input restrictions stipulated for tractable inference (Doyle and Patil 1991).

The most common response to complexity has been to use various speedup techniques and approximations in the hope of getting reasonable behaviour. AI has developed a very powerful armoury of methods for reducing the computational cost of decision making, including heuristic evaluation functions, pruning techniques, sampling methods, problem decomposition, hierarchical abstraction, compilation, and the application of metalevel control. Although some of these methods can retain guarantees of optimality and are effective for moderately large problems that are well structured, it is inevitable that intelligent agents will be unable to act rationally in all circumstances. This observation has been a commonplace since the very beginning of AI. Yet systems that select suboptimal actions fall outside calculative rationality per se, and we need a better theory to understand them.

2.5 Metalevel Rationality

Metalevel rationality, also called Type II rationality by Good (1971), is based on the idea of finding an optimal tradeoff between computational costs and decision quality. Although Good never made his concept of Type II rationality very precise—he defines it as "the maximization of expected utility *taking into account deliberation costs*"—it is clear that the aim was to take advantage of some sort of *metalevel architecture* to implement this tradeoff. Metalevel architecture is a design philosophy for intelligent agents that divides the agent program into two (or more) notional parts. The *object level* carries out computations concerned with the application domain—for example, projecting the results of physical actions, computing the utility of certain states, and so on. The *metalevel* is a second decision-making process whose application domain consists of the object-level computations themselves and the computational objects and states that they affect. Metareasoning

has a long history in AI, going back at least to the early 1970s (see Russell and Wefald 1991a, for historical details). One can also view selective search methods and pruning strategies as embodying metalevel expertise concerning the desirability of pursuing particular object-level search operations.

The theory of *rational metareasoning* formalizes Good's intuition that the metalevel can "do the right thinking." The basic idea is that object-level computations are actions with costs (the passage of time) and benefits (improvements in decision quality). A rational metalevel selects computations according to their expected utility. Rational metareasoning has as a precursor the theory of *information value* (Howard 1966)—the notion that one can calculate the decision-theoretic value of acquiring an additional piece of information by simulating the decision process that would be followed given each possible outcome of the information request, thereby estimating the expected improvement in decision quality averaged over those outcomes. The application to computational processes, by analogy to information-gathering, seems to have originated with Matheson (1968). In AI, Horvitz (1987, 1989), Breese and Fehling (1990), and Russell and Wefald (1989, 1991a,b) all showed how the idea of value of computation could solve the basic problems of real-time decision making.

Perhaps the simplest form of metareasoning occurs when the object level is viewed by the metalevel as a black-box *anytime* (Dean and Boddy 1988) or *flexible* (Horvitz 1987) algorithm, i.e., an algorithm whose decision quality depends on the amount of time allocated to computation. This dependency can be represented by a *performance profile* and the metalevel simply finds the optimal tradeoff between decision quality and the cost of time (Simon 1955). More complex problems arise if one wishes to build complex real-time systems from anytime components. First, one has to ensure the *interruptibility* of the composed system—that is, to ensure that the system as a whole can respond robustly to immediate demands for output. The solution is to interleave the execution of all the components, allocating time to each component so that the total time for each complete iterative improvement cycle of the system doubles at each iteration. In this way, we can construct a complex system that can handle arbitrary and unexpected real-time demands just as if it knew the exact time available in advance, with just a small ($\leq$4) constant factor penalty in speed (Russell and Zilberstein 1991). Second, one has to allocate the available computation optimally among the components to maximize the total output quality. Although this is NP-hard for the general case, it can be solved in time linear in program size when the call graph of the components is tree-structured (Zilberstein and Russell 1996). Although these results are derived in the simple context of anytime algorithms with well-defined performance profiles, they point to the possibility of more general schemes for allocation of computational resources in complex systems.

The situation gets more interesting when the metalevel can go inside the object level and direct its activities, rather than just switching it on and off. The work done with Eric Wefald looked in particular at search algorithms, in which the object-level computations extend projections of the results of various courses of actions further into the future. For example, in chess programs, each object-level

computation expands a leaf node of the game tree and advances the clock; it is an action in the so-called *joint-state Markov decision process*, whose state space is the Cartesian product of the object-level state space (which includes time) and the metalevel state space of computational states—in this case, partially generated game trees. The actions available are to expand a leaf of the game tree or to terminate search and make a move on the board. It is possible to derive a greedy or *myopic* approximation to the value of each possible computation and thereby to control search effectively. This method was implemented for two-player games, two-player games with chance nodes, and single-agent search. In each case, the same general metareasoning scheme resulted in efficiency improvements of roughly an order of magnitude over traditional, highly-engineered algorithms (Russell and Wefald 1991a).

An independent thread of research on metalevel control began with work by Kocsis and Szepesvari (2006) on the UCT algorithm, which operates in the context of Monte Carlo tree search (MCTS) algorithms. In MCTS, each computation takes the form of a simulation of a randomized sequence of actions leading from a leaf of the current tree to a terminal state. UCT is a metalevel heuristic for selecting a leaf from which to conduct the next simulation, and has contributed to dramatic improvements in Go-playing algorithms over the last few years. It views the metalevel decision problem as a multi-armed bandit problem (Berry and Fristedt 1985) and applies an asymptotically near-optimal bandit decision rule recursively to make a choice of which computation to do next. The application of bandit methods to metalevel control seems quite natural, because a bandit problem involves deciding where to do the next "experiment" to find out how good each bandit arm is. Are bandit algorithms such as UCT approximate solutions to some particular case of the metalevel decision problem defined by Russell and Wefald? The answer, perhaps surprisingly, is no. The essential difference is that, in bandit problems, every trial involves executing a real object-level action with real costs, whereas in the metareasoning problem the trials are *simulations* whose cost is usually *independent* of the utility of the action being simulated. Hence UCT applies bandit algorithms to problems that are not bandit problems. A careful analysis (Hay et al. 2012) shows that metalevel problems in their simplest form are isomorphic to *selection problems*, a class of statistical decision problems studied since the 1950s in quality control and other areas. Hay et al. develop a rigorous mathematical framework for metalevel problems, showing that, for some cases, hard upper bounds can be established for the number of computations undertaken by an optimal metalevel policy, while, for other cases, the optimal policy may (with vanishingly small probability) continue computing long past the point where the cost of computation exceeds the value of the object-level problem.

Achieving accurate metalevel control remains a difficult open problem in the general case. Myopic strategies—considering just one computation at a time—can fail in cases where multiple computations are required to have any chance of altering the agent's current preferred action. Obviously, the problem of optimal selection of computation *sequences* is at least as intractable as the underlying object-level problem. One possible approach could be to apply metalevel reinforcement learning,

especially as the "reward function" for computation—that is, the improvement in decision quality—is easily available to the metalevel post hoc. It seems plausible that the human brain has such a capacity, since its hardware is unlikely to have a method of deriving clever new algorithms for new classes of decision problems. Indeed, there is a sense in which *algorithms are not a necessary part of AI systems*. Instead, one can imagine a general, adaptive process of rationally guided computation interacting with properties of the environment to produce more and more efficient decision making.

Although rational metareasoning seems to be a useful tool in coping with complexity, the concept of metalevel rationality as a formal framework for resource-bounded agents does not seem to hold water. The reason is that, since metareasoning is expensive, it cannot be carried out optimally. Thus, while a metalevel-rational agent would be highly desirable (although not quite as desirable as a perfectly rational agent), it does not usually exist. The history of object-level rationality has repeated itself at the metalevel: perfect rationality at the metalevel is unattainable and calculative rationality at the metalevel is useless. Therefore, a time/optimality tradeoff has to be made for metalevel computations, as for example with the myopic approximation mentioned above. Within the framework of metalevel rationality, however, there is no way to identify the appropriate tradeoff of time for metalevel decision quality. Any attempt to do so via a metametalevel simply results in a conceptual regress. Furthermore, it is entirely possible that in some environments, the most effective agent design will do no metareasoning at all, but will simply respond to circumstances. These considerations suggest that the right approach is to step outside the agent, as it were; to refrain from micromanaging the individual decisions made by the agent. This is the approach taken in bounded optimality.

2.6 Bounded Optimality

The difficulties with perfect rationality and metalevel rationality arise from the imposition of optimality constraints on *actions* or *computations*, neither of which the agent designer directly controls. The basic problem is that not all agent functions are *feasible* (Russell and Subramanian 1995) on a given machine M; the feasible functions are those implemented by some program for M. Thus, the optimization over functions in Eq. (2.1) is meaningless. It may be pointed out that not all agent functions are computable, but feasibility is in fact much stricter than computability, because it relates the operation of a program on a formal machine model with finite speed to the actual temporal behaviour generated by the agent.

Given this view, one is led immediately to the idea that optimal feasible behaviour is an interesting notion, and to the idea of finding the program that generates it. P_4, bounded optimality, is exhibited by a program l_{opt} that satisfies

$$l_{\text{opt}} = \text{argmax}_{l \in \mathscr{L}_M} V(Agent(l, M), \mathbf{E}, U). \tag{2.2}$$

Certainly, one would be happy to have l_{opt}, which is as intelligent as possible given the computational resources and structural constraints of the machine M. Certainly, bounded optimal programs exist, by definition. And the research agenda appears to be very interesting, even though it is difficult.

In AI, the idea of bounded optimality floated around among several discussion groups interested in resource-bounded rationality in the late 1980s, particularly those at Rockwell (organized by Michael Fehling) and Stanford (organized by Michael Bratman). The term itself seems to have been originated by Horvitz (1989), who defined it informally as "the optimization of computational utility given a set of assumptions about expected problems and constraints on resources."

Similar ideas also emerged in game theory, where there has been a shift from consideration of optimal decisions in games to a consideration of optimal decision-making programs. This leads to different results because it limits the ability of each agent to do unlimited simulation of the other, who is also doing unlimited simulation of the first, and so on. Depending on the precise machine limitations chosen, it is possible to prove, for example, that the iterated Prisoner's Dilemma has cooperative equilibria (Megiddo and Wigderson 1986; Papadimitriou and Yannakakis 1994; Tennenholtz 2004), which is not the case for arbitrary strategies.

Philosophy has also seen a gradual evolution in the definition of rationality. There has been a shift from consideration of *act utilitarianism*—the rationality of individual acts—to *rule utilitarianism*, or the rationality of general policies for acting. The requirement that policies be feasible for limited agents was discussed extensively by Cherniak (1986) and Harman (1983). A philosophical proposal generally consistent with the notion of bounded optimality can be found in the "Moral First Aid Manual" (Dennett 1988). Dennett explicitly discusses the idea of reaching an optimum within the space of feasible decision procedures, using as an example the Ph.D. admissions procedure of a philosophy department. He points out that the bounded optimal admissions procedure may be somewhat messy and may have no obvious hallmark of "optimality"—in fact, the admissions committee may continue to tinker with it since bounded optimal systems may have no way to recognize their own bounded optimality.

My work with Devika Subramanian placed the general idea of bounded optimality in a formal setting and derived the first rigorous results on bounded optimal programs (Russell and Subramanian 1995). This required setting up completely specified relationships among agents, programs, machines, environments, and time. We found this to be a very valuable exercise in itself. For example, the informal notions of "real-time environments" and "deadlines" ended up with definitions rather different than those we had initially imagined. From this foundation, a very simple machine architecture was investigated in which the program consists of a collection of decision procedures with fixed execution time and decision quality. In a "stochastic deadline" environment, it turns out that the utility attained by running several procedures in sequence until interrupted is often higher than that attainable by any single decision procedure. That is, it is often better first to prepare a "quick and dirty" answer before embarking on more involved calculations in case the latter do not finish in time. In an entirely separate line of inquiry, Livnat and

Pippenger (2006) show that, under a bound on the total number of gates in a circuit-based agent, the bounded optimal configuration may, for some task environments, involve two or more separate circuits that compete for control of the effectors and, in essence, pursue separate goals.

The interesting aspect of these results, beyond their value as a demonstration of nontrivial proofs of bounded optimality, is that they exhibit in a simple way what I believe to be a major feature of bounded optimal agents: the fact that the pressure towards optimality within a finite machine results in more complex program structures. Intuitively, efficient decision-making in a complex environment requires a software architecture that offers a wide variety of possible computational options, so that in most situations the agent has at least some computations available that provide a significant increase in decision quality.

One objection to the basic model of bounded optimality outlined above is that solutions are not *robust* with respect to small variations in the environment or the machine. This in turn would lead to difficulties in analyzing complex system designs. Theoretical computer science faced the same problem in describing the running time of algorithms, because counting steps and describing instruction sets exactly gives the same kind of fragile results on optimal algorithms. The $O()$ notation was developed to provide a way to describe complexity that is independent of machine speeds and implementation details and that supports the cumulative development of complexity results. The corresponding notion for agents is asymptotic bounded optimality (ABO) (Russell and Subramanian 1995). As with classical complexity, we can define both average-case and worst-case ABO, where "case" here means the environment. For example, worst-case ABO is defined as follows:

Worst-case asymptotic bounded optimality
an agent program l is timewise (or spacewise) worst-case ABO in $\mathbf{E}$ *on* M *iff*

$$\exists k, n_0 \; \forall l', n \;\; n > n_0 \Rightarrow V^*(Agent(l, kM), \mathbf{E}, U, n) \geq V^*(Agent(l', M), \mathbf{E}, U, n)$$

where kM denotes a version of M speeded up by a factor k (or with k times more memory) and $V^*(f, \mathbf{E}, U, n)$ is the minimum value of $V(f, E, U)$ for all E in $\mathbf{E}$ *of complexity n.*

In English, this means that the program is basically along the right lines if it just needs a faster (larger) machine to have worst-case behaviour as good as that of any other program in all environments.

Another possible objection to the idea of bounded optimality is that it simply shifts the intractable computational burden of metalevel rationality from the agent's metalevel to the designer's object level. Surely, one might argue, the designer now has to solve offline all the metalevel optimization problems that were intractable when online. This argument is not without merit—indeed, it would be surprising

if the agent design problem turns out to be easy. There is however, a significant difference between the two problems, in that the agent designer is presumably creating an agent for an entire class of environments, whereas the putative metalevel agent is working in a specific environment. That this can make the problem *easier* for the designer can be seen by considering the example of sorting algorithms. It may be very difficult indeed to sort a list of a trillion elements, but it is relatively easy to design an asymptotically optimal algorithm for sorting. In fact, the difficulties of the two tasks are unrelated. The unrelatedness would still hold for BO as well as ABO design, but the ABO definitions make it a good deal clearer.

It can be shown easily that worst-case ABO is a generalization of asymptotically optimal algorithms, simply by constructing a "classical environment" in which classical algorithms operate and in which the utility of the algorithm's behaviour is a decreasing positive function of runtime if the output is correct and zero otherwise. Agents in more general environments may need to trade off output quality for time, generate multiple outputs over time, and so on. As an illustration of how ABO is a useful abstraction, one can show that under certain restrictions one can construct *universal* ABO programs that are ABO for any time variation in the utility function, using the doubling construction from Russell and Zilberstein (1991). Further directions for bounded optimality research are discussed below.

2.7 What Is to Be Done?

The 1997 version of this paper described two agendas for research: one agenda extending the tradition of calculative rationality and another dealing with metareasoning and bounded optimality.

2.7.1 Improving the Calculative Toolbox

The traditional agenda took as its starting point the kind of agent could be built using the components available at that time: a dynamic Bayesian network to model a partially observable, stochastic environment; parametric learning algorithms to improve the model; a particle filtering algorithm to keep track of the environment state; reinforcement learning to improve the decision function given the state estimate. Such an architecture "breaks" in several ways when faced with the complexity of real-world environments (Russell 1998):

1. Dynamic Bayesian networks are not expressive enough to handle environments with many related objects and uncertainty about the existence and identity of objects; a more expressive language—essentially a unification of probability and first-order logic—is required.

2. A flat space of primitive action choices, especially when coupled with a greedy decision function based on reinforcement learning, cannot handle environments where the relevant time scales are much longer than the duration of a single primitive action. (For example, a human lifetime involves tens of trillions of primitive muscle activation cycles.) The agent architecture must support hierarchical representations of behaviour, including high-level actions over long time scales.
3. Attempting to learn a value function accurate enough to support a greedy one-step decision procedure is unlikely to work; the decision function must support model-based lookahead over a hierarchical action model.

On this traditional agenda, a great deal of progress has occurred. For the first item, there are declarative (Milch et al. 2005) and procedural (Pfeffer 2001; Goodman et al. 2008) *probabilistic programming languages* that have the required expressive power. For the second item, a theory of hierarchical reinforcement learning has been developed (Sutton et al. 1999; Parr and Russell 1998). The theory can be applied to agent architectures defined by arbitrary *partial programs*—that is, agent programs in which the choice of action at any point may be left unspecified (Andre and Russell 2002; Marthi et al. 2005). The hierarchical reinforcement learning process converges in the limit to the optimal completion of the agent program, allowing the effective learning of complex behaviours that cover relatively long time scales. For the third item, lookahead over long time scales, a satisfactory semantics has been defined for high-level actions, at least in the deterministic setting, enabling model-based lookahead at multiple levels of abstraction (Marthi et al. 2008).

These are promising steps, but many problems remain unsolved. From a practical point of view, inference algorithms for expressive probabilistic languages remain far too slow, although this is the subject of intense study at present in many research groups around the world. Furthermore, algorithms capable of learning new model structures in such languages are in their infancy. The same is true for algorithms that construct new hierarchical behaviours from more primitive actions: it seems inevitable that intelligent systems will need high-level actions, but as yet we do not know how to create new ones automatically. Finally, there have been few efforts at integrating these new technologies into a single agent architecture. No doubt such an attempt will reveal new places where our ideas break and need to be replaced with better ones.

2.7.2 Optimizing Computational Behaviour

A pessimistic view of Eq. (2.2) is that it requires evaluating every possible program in order to find one that works best—hardly the most promising or original strategy for AI research. But in fact the problem has a good deal of structure and it is possible to prove bounded optimality results for reasonably general classes of machines and task environments.

Modular design using a hierarchy of components is commonly seen as the only way to build reliable complex systems. The components fulfill certain behavioural specifications and interact in well-defined ways. To produce a composite bounded-optimal design, the optimization problem involves allocating execution time to components (Zilberstein and Russell 1996) or arranging the order of execution of the components (Russell and Subramanian 1995) to maximize overall performance. As illustrated earlier in the discussion of universal ABO algorithms, the techniques for optimizing temporal behaviour are largely orthogonal to the *content* of the system components, which can therefore be optimized separately. Consider, for example, a composite system that uses an anytime inference algorithm over a Bayesian network as one of its components. If a learning algorithm improves the accuracy of the Bayesian network, the performance profile of the inference component will improve, which will result in a reallocation of execution time that is guaranteed to improve overall system performance. Thus, techniques such as the doubling construction and the time allocation algorithm of Zilberstein and Russell (1996) can be seen as domain-independent tools for agent design. They enable bounded optimality results that do not depend on the specific temporal aspects of the environment class. As a simple example, we might prove that a certain chess program design is ABO for all time controls ranging from blitz to full tournament play.

The results obtained so far for optimal time allocation have assumed a static, offline optimization process with predictable component performance profiles and fixed connections among components. One can imagine far more subtle designs in which individual components must deal with unexpectedly slow or fast progress in computations and with changing needs for information from other components. This might involve exchanging computational resources among components, establishing new interfaces, and so on. This is more reminiscent of a computational market, as envisaged by Wellman (1994), than of the classical subroutine hierarchies, and would offer a useful additional level of abstraction in system design.

2.7.3 Learning and Bounded Optimality

In addition to combinatorial optimization of the structure and temporal behaviour of an agent, we can also use learning methods to improve the design:

- The *content* of an agent's knowledge base can of course be improved by inductive learning. Russell and Subramanian (1995) show that approximately bounded optimal designs can be guaranteed with high probability if each component is learned in such a way that its output quality is close to optimal among all components of a given execution time. Results from statistical learning theory, particularly in the agnostic learning and empirical risk minimization models (Kearns et al. 1992; Vapnik 2000), can provide learning methods—such as support vector machines—with the required properties. The key additional

step is to analyze the way in which slight imperfection in each component carries through to slight imperfection in the whole agent.

- *Reinforcement learning* can be used to learn value information such as utility functions, and several kinds of ϵ-δ convergence guarantees have been established for such algorithms. Applied in the right way to the metalevel decision problem, a reinforcement learning process can be shown to converge to a bounded-optimal configuration of the overall agent.
- *Compilation* methods such as explanation-based learning can be used to transform an agent's representations to allow faster decision making. Several agent architectures including SOAR (Laird et al. 1986) use compilation to speed up all forms of problem solving. Some nontrivial results on convergence have been obtained by Tadepalli (1991), based on the observation that after a given amount of experience, novel problems for which no solution has been stored should be encountered only infrequently.

Presumably, an agent architecture can incorporate all these learning mechanisms. One of the issues to be faced by bounded optimality research is how to prove convergence results when several adaptation and optimization mechanisms are operating simultaneously.

2.7.4 Offline and Online Mechanisms

One can distinguish between *offline* and *online* mechanisms for constructing bounded-optimal agents. An offline construction mechanism is not itself part of the agent and is not the subject of bounded optimality constraints. Let C be an offline mechanism designed for a class of environments $\mathbf{E}$. Then a typical theorem will say that C operates in a specific environment $E \in \mathbf{E}$ and returns an agent design that is ABO (say) for E—that is, an environment-specific agent.

In the online case, the mechanism C is considered part of the agent. Then a typical theorem will say that the agent is ABO for all $E \in \mathbf{E}$. If the performance measure used is indifferent to the transient cost of the adaptation or optimization mechanism, the two types of theorems are essentially the same. On the other hand, if the cost cannot be ignored—for example, if an agent that learns quickly is to be preferred to an agent that reaches the same level of performance but learns more slowly—then the analysis becomes more difficult. It may become necessary to define asymptotic equivalence for "experience efficiency" in order to obtain robust results, as is done in computational learning theory.

It is worth noting that one can easily prove the value of "lifelong learning" in the ABO framework. An agent that devotes a constant fraction of its computational resources to learning-while-doing cannot do worse, in the ABO sense, than an agent that ceases learning after some point. If some improvement is still possible, the lifelong learning agent will always be preferred.

2.7.4.1 Fixed and Variable Computation Costs

Another dimension of design space emerges when one considers the computational cost of the "variable part" of the agent design. The design problem is simplified considerably when the cost is fixed. Consider again the task of metalevel reinforcement learning, and to make things concrete let the metalevel decision be made by a Q function mapping from computational state and action to value. Suppose further that the Q function is to be represented by a neural net. If the topology of the neural net is fixed, then all Q functions in the space have the same execution time. Consequently, the optimality criterion used by the standard Q-learning process coincides with bounded optimality, and the equilibrium reached will be a bounded-optimal configuration.[1] On the other hand, if the topology of the network is subject to alteration as the design space is explored, then the execution time of the different Q-functions varies. In this case, the standard Q-learning process will not necessarily converge to a bounded-optimal configuration; typically, it will tend to build larger and larger (and therefore more and more computationally expensive) networks to obtain a more accurate approximation to the true Q-function. A different adaptation mechanism must be found that takes into account the passage of time and its effect on utility.

Whatever the solution to this problem turns out to be, the important point is that the notion of bounded optimality helps to distinguish adaptation mechanisms that will result in good performance from those that will not. Adaptation mechanisms derived from calculative rationality will fail in the more realistic setting where an agent cannot afford to aim for perfection.

2.7.5 Looking Further Ahead

The discussion so far has been limited to fairly sedate forms of agent architecture in which the scope for adaptation is circumscribed to particular functional aspects such as metalevel Q functions. However, an agent must in general deal with an environment that is far more complex than itself and that exhibits variation over time at all levels of granularity. Limits on the size of the agent's memory may imply that almost complete revision of the agent's mental structure is needed to achieve high performance. (For example, songbirds grow their brains substantially during the singing season and shrink them again when the season is over.) Such situations may engender a rethinking of some of our notions of agent architecture and optimality, and suggest a view of agent programs as dynamical systems with various amounts of compiled and uncompiled knowledge and internal processes of inductive learning, forgetting, and compilation.

[1] A similar observation was made by Horvitz and Breese (1990) for cases where the object level is so restricted that the metalevel decision problem can be solved in constant time.

If a true science of intelligent agent design is to emerge, it will have to operate in the framework of bounded optimality. One general approach—discernible in the examples given earlier—is to divide up the space of agent designs into "architectural classes" such that in each class the structural variation is sufficiently limited. Then ABO results can be obtained either by analytical optimization within the class or by showing that an empirical adaptation process results in an approximately ABO design. Once this is done, it should be possible to compare architecture classes directly, perhaps to establish asymptotic dominance of one class over another. For example, it might be the case that the inclusion of an appropriate "macro-operator formation" or "greedy metareasoning" capability in a given architecture will result in an improvement in behaviour in the limit of very complex environments—that is, one cannot compensate for the exclusion of the capability by increasing the machine speed by a constant factor. Moreover, within any particular architectural class it is clear that faster processors and larger memories lead to dominance. A central tool in such work will be the use of "no-cost" results where, for example, the allocation of a constant fraction of computational resources to learning or metareasoning can do no harm to an agent's ABO prospects.

Getting all these architectural devices to work together smoothly is an important unsolved problem in AI and must be addressed before we can make progress on understanding bounded optimality within these more complex architectural classes. If the notion of "architectural device" can be made sufficiently concrete, then AI may eventually develop a *grammar* for agent designs, describing the devices and their interrelations. As the grammar develops, so should the accompanying ABO dominance results.

2.8 Summary

I have outlined some directions for formally grounded AI research based on bounded optimality as the desired property of AI systems. This perspective on AI seems to be a logical consequence of the inevitable philosophical "move" from optimization over actions or computations to optimization over programs. I have suggested that such an approach should allow synergy between theoretical and practical AI research of a kind not afforded by other formal frameworks. In the same vein, I believe it is a satisfactory formal counterpart of the informal goal of creating intelligence. In particular, it is entirely consistent with our intuitions about the need for complex structure in real intelligent agents, the importance of the resource limitations faced by relatively tiny minds in large worlds, and the operation of evolution as a design optimization process. One can also argue that bounded optimality research is likely to satisfy better the needs of those who wish to emulate human intelligence, because it takes into account the limitations on computational resources that are presumably an important factor in the way human minds are structured and in the behaviour that results.

Bounded optimality and its asymptotic version are, of course, nothing but formally defined properties that one may want systems to satisfy. It is too early to tell whether ABO will do the same kind of work for AI that asymptotic complexity has done for theoretical computer science. Creativity in design is still the prerogative of AI researchers. It may, however be possible to systematize the design process somewhat and to automate the process of adapting a system to its computational resources and the demands of the environment. The concept of bounded optimality provides a way to make sure the adaptation process is "correct."

My hope is that with these kinds of investigations, it will eventually be possible to develop the conceptual and mathematical tools to answer some basic questions about intelligence. For example, *why* do complex intelligent systems (appear to) have declarative knowledge structures over which they reason explicitly? This has been a fundamental assumption that distinguishes AI from other disciplines for agent design, yet the answer is still unknown. Indeed, Rod Brooks, Hubert Dreyfus, and others flatly deny the assumption. What is clear is that it will need *something like* a theory of bounded optimal agent design to answer this question.

Most of the agent design features that I have discussed here, including the use of declarative knowledge, have been conceived within the standard methodology of "first build calculatively rational agents and then speed them up." Yet one can legitimately doubt that this methodology will enable the AI community to discover all the design features needed for general intelligence. The reason is that no conceivable computer will ever be remotely close to approximating perfect rationality for even moderately complex environments. It may well be the case, therefore, that agents based on approximations to calculatively rational designs are *not even close* to achieving the level of performance that is potentially achievable given the underlying computational resources. For this reason, I believe it is imperative not to dismiss ideas for agent designs that do not seem at first glance to fit into the "classical" calculatively rational framework.

Acknowledgements An earlier version of this paper appeared in the journal *Artificial Intelligence*, published by Elsevier. That paper drew on previous work with Eric Wefald and Devika Subramanian. More recent results were obtained with Nick Hay. Thanks also to Michael Wellman, Michael Fehling, Michael Genesereth, Russ Greiner, Eric Horvitz, Henry Kautz, Daphne Koller, Bart Selman, and Daishi Harada for many stimulating discussions topic of bounded rationality. The research was supported by NSF grants IRI-8903146, IRI-9211512 and IRI-9058427, and by a UK SERC Visiting Fellowship. The author is supported by the *Chaire Blaise Pascal*, funded by the l'État et la Région Île de France and administered by the Fondation de l'École Normale Supérieure.

References

Agre, P. E., & Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, Milan (pp. 268–272). Morgan Kaufmann.

Andre, D., & Russell, S. J. (2002) State abstraction for programmable reinforcement learning agents. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02)*, Edmonton (pp. 119–125). AAAI Press.

Bellman, R. E. (1957). *Dynamic programming*. Princeton: Princeton University Press.

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments*. London: Chapman and Hall.

Breese, J. S., & Fehling, M. R. (1990). Control of problem-solving: Principles and architecture. In R. D. Shachter, T. Levitt, L. Kanal, & J. Lemmer (Eds.), *Uncertainty in artificial intelligence 4*. Amsterdam/London/New York: Elsevier/North-Holland.

Brooks, R. A. (1989). Engineering approach to building complete, intelligent beings. *Proceedings of the SPIE—The International Society for Optical Engineering, 1002*, 618–625.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: University of Chicago Press.

Cherniak, C. (1986). *Minimal rationality*. Cambridge: MIT.

Dean, T., & Boddy, M. (1988) An analysis of time-dependent planning. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-88)*, St. Paul (pp. 49–54). Morgan Kaufmann.

Dean, T., Aloimonos, J., & Allen, J. F. (1995). *Artificial intelligence: Theory and practice*. Redwood City: Benjamin/Cummings.

Dennett, D. C. (1988). The moral first aid manual. In S. McMurrin (Ed.), *Tanner lectures on human values* (Vol. 7, pp. 121–147). University of Utah Press and Cambridge University Press.

Doyle, J., & Patil, R. (1991). Two theses of knowledge representation: Language restrictions, taxonomic classification, and the utility of representation services. *Artificial Intelligence, 48*(3), 261–297

Good, I. J. (1971) Twenty-seven principles of rationality. In V. P. Godambe & D. A. Sprott (Eds.), *Foundations of statistical inference* (pp. 108–141). Toronto: Holt, Rinehart, Winston.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Proceedings of UAI-08*, Helsinki (pp. 220–229).

Harman, G. H. (1983). *Change in view: Principles of reasoning*. Cambridge: MIT.

Hay, N., Russell, S., Shimony, S. E., & Tolpin, D. (2012). Selecting computations: Theory and applications. In *Proceedings of UAI-12*, Catalina Island.

Horvitz, E. J. (1987). Problem-solving design: Reasoning about computational value, trade-offs, and resources. In *Proceedings of the Second Annual NASA Research Forum, NASA Ames Research Center*, Moffett Field, CA (pp. 26–43).

Horvitz, E. J. (1989). Reasoning about beliefs and actions under computational resource constraints. In L. N. Kanal, T. S. Levitt, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence 3* (pp. 301–324). Amsterdam/London/New York: Elsevier/North-Holland.

Horvitz, E. J., & Breese, J. S. (1990). Ideal partition of resources for metareasoning (Technical report KSL-90-26), Knowledge Systems Laboratory, Stanford University, Stanford.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, SSC-2*, 22–26.

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin/New York: Springer.

Kearns, M., Schapire, R. E., & Sellie, L. (1992). Toward efficient agnostic learning. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (COLT-92)*, Pittsburgh. ACM.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Kocsis, L., & Szepesvari, C. (2006). Bandit-based Monte-Carlo planning. In *Proceedings of ECML-06*, Berlin.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission, 1*(1), 1–7.

Koopmans, T. C. (1972). Representation of preference orderings over time. In C.B. McGuire & R. Radner (Eds.), *Decision and organization*. Amsterdam/London/New York: Elsevier/North-Holland.

Kumar, P. R., & Varaiya, P. (1986). *Stochastic systems: Estimation, identification, and adaptive control*. Upper Saddle River: Prentice-Hall.

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning, 1*, 11–46.

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence, 30*(1), 81–108.

Livnat, A., & Pippenger, N. (2006). An optimal brain can be composed of conflicting agents. *Proceedings of the National Academy of Sciences of the United States of America 103*(9), 3198–3202.

Marthi, B., Russell, S., Latham, D., & Guestrin, C. (2005). Concurrent hierarchical reinforcement learning. In *Proceedings of IJCAI-05*, Edinburgh.

Marthi, B., Russell, S. J., & Wolfe, J. (2008). Angelic hierarchical planning: Optimal and online algorithms. In *Proceedings of ICAPS-08*, Sydney.

Matheson, J. E. (1968). The economic value of analysis and computation. *IEEE Transactions on Systems Science and Cybernetics, SSC-4*(3), 325–332.

Megiddo, N., & Wigderson, A. (1986). On play by means of computing machines. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference (TARK-86), IBM and AAAI*, Monterey (pp. 259–274). Morgan Kaufmann.

Milch, B., Marthi, B., Sontag, D., Russell, S. J., Ong, D., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *Proceedings of IJCAI-05*, Edinburgh.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*(1), 82–127.

Nilsson, N. J. (1991). Logic and artificial intelligence. *Artificial Intelligence, 47*(1–3), 31–56

Papadimitriou, C. H., & Yannakakis, M. (1994). On complexity as bounded rationality. In *Symposium on Theory of Computation (STOC-94)*, Montreal.

Parr, R., & Russell, S. J. (1998). Reinforcement learning with hierarchies of machines. In M. I. Jordan, M. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10*. Cambridge: MIT.

Pfeffer, A. (2001). IBAL: A probabilistic rational programming language. In *Proceedings of IJCAI-01*, Seattle (pp. 733–740).

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence, 94*, 57–77.

Russell, S. J. (1998). Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual ACM Workshop on Computational Learning Theory (COLT-98)*, Madison (pp. 101–103). ACM.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Upper Saddle River: Prentice-Hall.

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research, 3*, 575–609.

Russell, S. J., & Wefald, E. H. (1989). On optimal game-tree search using rational meta-reasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI-89)*, Detroit (pp. 334–340). Morgan Kaufmann.

Russell, S. J., & Wefald, E. H. (1991a). *Do the right thing: Studies in limited rationality*. Cambridge: MIT.

Russell, S. J., & Wefald, E. H. (1991b). Principles of metareasoning. *Artificial Intelligence 49*(1–3), 361–395.

Russell, S. J., & Zilberstein, S. (1991). Composing real-time systems. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, Sydney. Morgan Kaufmann.

Shoham, Y., & Leyton-Brown, K. (2009). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge/New York: Cambridge University Press.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Simon, H. A. (1958). Rational choice and the structure of the environment. In *Models of bounded rationality* (Vol. 2). Cambridge: MIT.

Solomonoff, R. J. (1964). A formal theory of inductive inference. *Information and Control, 7*, 1–22, 224–254.

Srivastava, S., Russell, S., Ruan, P., & Cheng, X. (2014). First-order open-universe POMDPs. In *Proceedings of UAI-14*, Quebec City.

Sutton, R., Precup, D., & Singh, S. P. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence, 112*, 181–211.

Tadepalli, P. (1991). A formalization of explanation-based macro-operator learning. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence (IJCAI-91)*, Sydney (pp. 616–622). Morgan Kaufmann.

Tennenholtz, M. (2004). Program equilibrium. *Games and Economic Behavior, 49*(2), 363–373.

Vapnik, V. (2000). *The nature of statistical learning theory*. Berlin/New York: Springer.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior* (1st ed.). Princeton: Princeton University Press.

Wellman, M. P. (1994). A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research, 1*(1), 1–23.

Wellman, M. P., & Doyle, J. (1991). Preferential semantics for goals. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, Anaheim (Vol. 2, pp. 698–703). AAAI Press.

Zilberstein, S., & Russell, S. J. (1996). Optimal composition of real-time systems. *Artificial Intelligence 83*, 181–213.

Chapter 3
Computation and Multiple Realizability

Marcin Miłkowski

Abstract Multiple realizability (MR) is traditionally conceived of as *the* feature of computational systems, and has been used to argue for irreducibility of higher-level theories. I will show that there are several ways a computational system may be seen to display MR. These ways correspond to (at least) five ways one can conceive of the function of the physical computational system. However, they do not match common intuitions about MR. I show that MR is deeply interest-related, and for this reason, difficult to pin down exactly. I claim that MR is of little importance for defending computationalism, and argue that it should rather appeal to organizational invariance or substrate neutrality of computation, which are much more intuitive but cannot support strong antireductionist arguments.

Keywords Multiple realizability • Functionalism • Computationalism

I want to undermine the conviction that multiple realizability (MR) is particularly important in understanding the nature of computational systems. MR is held to be fundamental as far as it is considered indispensable in arguing for irreducibility of theories that appeal to the notion of computation. This is why this conviction is especially common among proponents of antireductionism who want to defend the autonomy of psychology (for example, (Block 1990, p. 146)), even if it's not so important for theorists of computability, if not completely alien to them.[1]

Recently, it was also argued that mechanistic theories of physical computation (Piccinini 2007, 2010) were not compatible with multiple realization (Haimovici 2013), which was supposed to show that mechanism is wrong. Indeed, in defending my mechanistic account of computation, I denied that multiple realization is an essential feature of computation, and that there are no facts of the matter that could

[1] I owe this observation Aaron Sloman.

M. Miłkowski (✉)
Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland
e-mail: marcin.milkowski@gmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_3

easily establish that a given computational capacity is actually multiply realized or not (Miłkowski 2013, Chap. 2). I want to develop this point in detail here and show that there are multiple ways of carving the computational capacity but whenever it is made precise, there are either too many examples of MR or scarcely any in cases that have been considered paradigmatic. For this reason, it does not seem to be an essential feature at all; I suggest that it can be easily replaced with another similar feature of organizational invariance or substrate neutrality.

The organization of the paper is as follows. First, I introduce the notion of MR, and show it is quite vague. I argue for additional restrictions on the notion to make it more precise. Then, I analyze one vivid example from the history of computing. In the following discussion, I distinguish MR from substrate neutrality and suggest that the latter is indeed more important. At the same time, substrate neutrality does not lend credibility to antireductionist arguments, and remains compatible with type-identity theory.

3.1 Multiple Realization Introduced, Criticized, and Made More Precise

The argument from multiple realizability to irreducibility is usually attributed to Jerry Fodor (1974) and to Hilary Putnam (1975). Elliot Sober summarizes the argument in the following way:

1. Higher-level sciences describe properties that are multiply realizable and that provide good explanations.
2. If a property described in a higher-level science is multiply realizable at a lower level, then the lower-level science will not be able to explain, or will explain only feebly, the phenomena that the higher-level science explains well.
3. If higher-level sciences provide good explanations of phenomena that lower-level sciences cannot explain, or explain only feebly, then reductionism is false.

Reductionism is false. (Sober 1999, p. 558)

However, it needs to be noted that neither Fodor nor Putnam uses the term "multiple realization" or its cognates. In the subsequent discussion, it was used mostly informally, usually without a definition but with the help of Fodor's example: money. Money can be realized as coins, banknotes, credit cards, sea shells, and what not. There is simply no physical kind that encompasses all realizations of money, Fodor claims, and the lower-level, physical description of the bearers of monetary value is bound to be wildly disjunctive (Fodor 1974). The same is supposed to happen with realizations of computer algorithms, but not simply because they are cultural entities, like money: it has been claimed that one can make a computer of silicon chips, but also of toilet paper or Swiss cheese, and they would realize the same algorithm and be different realizations. Simply speaking, a functional capacity is multiply realized if and only if its occurrence is owing to different realizing structures. But this initial definition remains fairly vague; it does not decide whether anything else but functional capacities is multiply realized; it also

does not set any standards on when functional capacities count as exactly of the same type, and what the standards for sameness of realizers are. It turns out that making the notion precise is extremely difficult. In addition, according to Keeley (2000), neuroethology uses both behavioral evidence and physiological evidence to justify its claims; and MR might be compatible with reductionism after all. But whether the reductionism is correct or not is not the question I am interested in this paper.

It's fair to say that after initial enthusiasm, philosophers have become much more skeptical about MR, and many argue that it need not be so frequent or essential. For example (Bechtel and Mundale 1999) argued that, contrary to appearances, different brains were at least sometimes viewed as exactly the *same* realizers of a given psychological capacity: Neuroscience frequently uses animal brains as models of the human brain, and there would be no animal models for human brains if animal brains were actually so wildly different as antireductionists suppose. For example, the rat's brain is not really a different realization of dopamine-related capacities when it uses dopamine. Just like Bechtel and Mundale, Polger (2004, 2008) and Shapiro (2000, 2004) consequently argued that there might be merely illusory MR in cases where the function has been described in a generic way and its realization quite specifically. Just as different colors of a corkscrew do not make them different physical realizations (after all, the color is irrelevant to the function of the corkscrew), not all specific details count when it comes to distinguishing different realizations. But even not all *causally* relevant detail is equally important: for example, for opening a wine bottle, one could easily use a corkscrew with five or four threads on their screws. There is a certain level of changes that we usually consider irrelevant (in the engineering contexts, one speaks of tolerance levels). What makes an amount of possible differences irrelevant? Are there any principled answers to this question? Before I go on, an important caveat is in place.

Some would deny that corkscrews are ever multiply realized. First, they are artifacts, and some deny functional status to artifacts or claim that artifacts only have functionality derived from the functions of their biological users (Davies 2001). Here, I will assume that derived functions are equally functional so as not to beg the question against the proponents of MR with respect to computation. Second, and more importantly, functional capacities of corkscrews and properties of realizations of their capacities to open bottles might be both ascribed simply to corkscrews, rather than parts of corkscrews. That leads to a general question. Is it possible for a functional capacity to be a property of the same entity as the properties of the realizers? A positive answer is the so-called 'standard', or 'flat' view on MR, whereas the 'dimensioned view' stresses that instances of properties of realizers are not the properties of the same entity, which introduces more 'dimensions' to the definition of MR (Aizawa and Gillett 2009; Gillett 2002). The dimensioned view has its roots in the functionalist tradition: Cummins' (1975) conditions for functional analysis require that functional capacities be ascribed to higher-level structures, and their realizing structures be on the lower-level; thus on the dimensioned view, MR

is an ontological inter-level relationship. The defenders of the flat view have argued, however, that the dimensioned view leads to absurdity (Polger and Shapiro 2008). My point in this paper is however largely independent from this debate (but for a reply, see (Gillett 2011)), and the main problem for MR of computation is essentially the same for both dimensioned and flat views. In what follows, I will still peruse the simple example of corkscrews; if you subscribe to the dimensioned view, there is always a paraphrase of my claims in terms of the operations and parts of the corkscrew rather in terms of the operation of the corkscrew itself.

Let me return to the core of the problem with MR. My initial definition is vague, and we need to have strict standards of sameness and difference to show that MR actually occurs, and these standards cannot beg the question against reductionism. But other definitions currently espoused by defenders of MR do not seem to fare any better. For example:

> (Multiple Realization) A property *G* is multiply realized if and only if (i) under condition *$*, an individual *s* has an instance of property *G* in virtue of the powers contributed by instances of properties/relations F_1 -F_n to *s*, or *s*'s constituents, but not vice versa; (ii) under condition *$** (which may or may not be identical to *$*), an individual *s** (which may or may not be identical to *s*) has an instance of a property *G* in virtue of the powers contributed by instances of properties/relations F^*_1 -F^*_m to *s** or *s**'s constituents, but not vice versa; (iii) F_1 -$F_n \neq F^*_1$-F^*_m and (iv), under conditions *$* and *$**, F_1 -F_n and F^*_1 -F^*_m are at the same scientific level of properties (Aizawa and Gillett 2009, p. 188).

According to (iii), we need to make sure that properties or relations are different in different realizers. But what are the standards, again? The ascription of function should be as fine-grained as the description of the function realization. However, the notion of function is inextricably linked with theoretical interests of the observer, thus MR is also heavily interest-dependent. (Note: even if there are proposals to make the function ascriptions as determinate as possible, they usually pertain only to the so-called notion of etiological notion of function, cf. (Price 2001); for MR, however, the relevant notion is systemic or dispositional, as defined by Cummins (1975), and this notion is interest-dependent.) It does not mean that there are absolutely no facts of the matter in function ascriptions, as ascriptions may fail to be true; but they are still interest-driven (Craver 2013).

So how could one decide when the functional capacity is the same, and when it is not? Is there any evidence one could use? For example, Shagrir noted that the evidence for the sameness of the psychological capacities in the case of neuroplasticity – or, more importantly, in case of realizing computational algorithms – is not really available for antireductionists, as all that they have is a set of "behavioral antecedents and consequences of subjects, and physical realizations of the mediating computational mechanisms" (Shagrir 1998). This kind of evidence simply cannot settle the question at all. For this reason, an important topic in the debate has become the question of how to evaluate claims for MR (Polger 2008; Shapiro 2008).

So how can we evaluate the claim that a given computational algorithm or a computational system has multiple realizations? Here are possible answers. I will argue that in interesting cases – the ones usually held to be paradigmatic – there's no MR, contrary to appearances.

3.2 When Are Computations Multiply Realized?

Let me pick a particularly vivid example of two compatible IBM computers, 709 and 7090, one with tubes, and another with transistors. It seems to be ideally suited to test the claim whether one can fruitfully talk of multiple realization in the case of computational systems. In addition, some authors have earlier considered two IBMs to be realizations of the same computational system (Wimsatt 2002), and I have argued earlier that it is not necessarily the case (Miłkowski 2013). (IBM 7090 was also featured in Dr. Strangelove, which should be a good enough reason to consider it worthy of philosophical attention.)

IBM 709 and IBM 7090 share the same logical diagram (as witnessed by their technical documentation) but they use different electronics. Transistors work faster than tubes, so they differ also with their speed of operation. They have obviously different footprint, as transistors are smaller: the newer version is 50 % smaller, and needs less ventilation because transistors need less cooling. Transistors also consume 70 % less power, while tubes are 6 times slower. There's no denying that these machine differ; they also share a lot (for more technical specifications and photos, see ("IBM Archives: 709 Data Processing System," 2003, "IBM Archives: 7090 Data Processing System," 2003)). But is there MR in the strict meaning of the term?

To make MR claims about computational precise, one has to explain what it is to have a computational function. I can enumerate at least five ways one can understand the computational capacity of a given computer. I will explain these in turn but before I do so, I need to briefly introduce my mechanistic framework that allows talking about physical – or implemented – computation. One of the most widely endorsed views in the philosophy of special sciences is neo-mechanism (Bechtel 2008; Craver 2007; Machamer et al. 2000). According to this view, to explain a phenomenon is to explain the underlying mechanism. Mechanistic explanation is causal explanation, and explaining a mechanism involves describing its causal structure. While mechanisms are defined in various ways by different authors, the core idea is that they are organized systems, comprising causally relevant component parts and operations (or activities) thereof. Components of the mechanism interact and their orchestrated operation contributes to the capacity of the mechanism.

A mechanism implements a computation just when the causal organization of the mechanism is such that the input and output information streams are causally linked and that this link, along with the specific structure of information processing, is completely described (the notion of information is not semantic in my account; for a similar treatment, see (Fresco 2014)). Importantly, the link can be cyclical and as complex as one could wish. Mechanistic constitutive explanation, usually used to explain physical computation in cases where we deem physical implementation important, includes *at least* three levels of the mechanism: a *constitutive* (−1) level, which is the lowest level in the given analysis; an *isolated* (0) level, at which the parts of the mechanism are specified along with their interactions (activities or

operations); and the *contextual* (+1) level, at which the function of the mechanism is seen in a broader context. These levels are not just levels of abstraction; they are levels of *composition* or organization.

The description of a *mechanistically adequate model* of computation comprises *two* parts: (1) an abstract specification of a computation, which should include all the variables causally relevant for the computation; (2) a complete blueprint of the mechanism on three levels of its organization. I call the first part the *formal model of the mechanism* and the second the *instantiation blueprint* of the mechanism (for a detailed study of how this framework is applied to physical and non-conventional computers, such as Physarum machines, see (Miłkowski 2014)).

After this brief introduction of terminology, I can enumerate five possible ways of understanding the capacity of the computational mechanism. First, the computers could share the same causal structure associated with their formal model and differ in their instantiation blueprint (let's call it *formal-model* MR). Now, how much do they have to differ? Arguably, one could say that a certain level of tolerance for physical changes is required; otherwise, simple physical wear and tear or replacement of worn parts would be enough to say that a computer is now a new realization of its older version (this is a new version of the ancient puzzle regarding the ship of Theseus). This is why only physical changes relevant to the execution of computation should matter; replacing a chip with a faster one would only make the speed of operation faster by a certain linear factor k (as is usually assumed in the theory of computational complexity), rather than mean that the computer has an altogether different substrate.

An emulation of the PowerPC computer on an Intel *x*86 architecture should qualify as formal-model MR, as they both share the formal model of the computation (in virtue of emulation on the second one) but the second machine has to include much more in its formal model, as the emulated machine is just virtual. So the Intel machine does not really have the *same* formal model; the formal model of PowerPC machine is rather a proper part of the complete formal model of the Intel *x*86 machine. In other words, these machines literally *share* the same formal model but they do not have exactly the same model.

Let me compare this case to what Bechtel and Mundale (1999) consider not to be an instance of MR. Usually, philosophers assumed that the functionality of the brain is multiply realized by various anatomies in different species. However, the human brain topographical map, created by Korbinian Brodman does rely on inter-species anatomical similarities, or homologies. In other words, anatomical similarity is used in function ascriptions. Actually, it seems that there are lower-level types that match higher-level types. The same consideration can be used in the case of my laptop: the kind of hardware it requires to function properly is specified with a certain level of tolerance. As long as memory chips, for example, match the specification of the motherboard, they can be used in my laptop without changing its functionality. But the similarity in question goes beyond the ability to replace parts; for example, one cannot replace the part of the human brain with a brain of a rodent and keep the same functionality. However, there are lower-level, anatomical types that still match the functionality of the brain in an orderly fashion, so that one can defend a

kind of type identity between both. It seems therefore cogent that as long as there is sufficient similarity that allows classifying tokens of the realizing structures as belonging to the same type we do not have any kind of MR. Actually, the reason for similarity between brains is homology, or shared ancestry. In the case of IBM 7090, it is similar to IBM 709 because the latter is its ancestor, and we should expect a lot of hardware similarities, even if we cannot substitute tubes with transistors directly (owing to different voltages and so on). For the case of emulation of a PowerPC on an Intel *x*86 CPU, the similarity of functionality in question is not caused by common ancestry but by the software run, so there are no such hardware similarities we might expect between both IBMs.

One could insist that vast physical differences between IBMs constitute different types. But there are also vast differences between how both IBMs operate, such as different speed of operation, which obviously makes difference for the use of computers, and by supposing that the functional capacity in question is what is specified just by the formal model of the mechanism, we decided to *ignore* such differences as irrelevant, which is an important and far-reaching assumption. Yet if we use low-grain distinctions for functional types, there is no *principled* reason to use fine-grain distinctions for structural types, beside the philosophical prejudice for MR. They are definitely software-compatible and could run the same FORTRAN programs.

A second way to understand the computational capacities of computers is to focus on the mathematical function that they compute, specified in terms of input and output values (this will be *mathematical function MR*). This is the level of computational equivalence usually presupposed in the computability theory: If one says that a computer C is Turing-machine-equivalent in terms of the set of C-computable functions, then one presupposes that they will produce the same output given the same input, whatever the way they compute it. Note however that this is not the level of grain usually presupposed in cognitive science when it speaks of computational algorithms realized by brains: in cognitive science, the way which a given output is produced is immensely important (Fodor 1968; Miłkowski 2011; Shagrir 1998). For example, reaction times are used as evidence to decide whether people use this or another algorithm (Meyer et al. 1988; Posner 2005), so algorithms are individuated with a finer grain in cognitive research.

Is there mathematical function MR for both IBMs? They definitely share the same mathematical functions (which follows immediately from the fact that they could use exactly the same software). Again, however, their hardware similarities make it possible to classify them as the same kind of machine – they aren't even realizing different computational architectures. The difference between tubes and transistors is just as irrelevant as the difference between four and five threads on my corkscrews.

So how would a cognitive scientist understand a computational capacity of a system? In this third version, the capacity would also include the features of the physical implementation, such as the speed of operation; this will be *instantiation blueprint MR*. Of course, there is again a certain level of tolerance, so that a mere replacement of one electronic part does not create another realization of the same

type. In the case of two IBMs, we have different speeds, as tubes are simply slower. Of course, the bigger footprint of the tube-based IBM is not so important here (just like cognitive scientists usually ignore the fact whether subjects are overweight or not) but there are different patterns of breakdown, which are also important in neuropsychology, as pathologies are used to make inferences about function (Glymour 1994). All in all, here even the capacity is different, so there is no chance for MR to occur.

Between the level of grain implied by mathematical input/output specifications and the one assumed in cognitive science, one can drive a wedge for another – fourth – specification. For example, any algorithm for sorting words alphabetically shares the same input/output relations but not the same sequence of steps, which is again specifiable in terms of input/output relations. So, one could understand a particular algorithm for sorting, such as QuickSort, as mathematical function that is decomposed into an ordered sequence of other more basic mathematical functions. The kind of MR in this case might be dubbed *decomposed mathematical function MR.* This is how, for example, Keeley seems to understand the algorithm underlying perceptual abilities of weakly electric fish (he cannot mean the whole formal model, as in the first meaning of the functional capacity above, because exact ways that the algorithm is realized are not known). Note however that we have reason to suppose that there is MR only when fish realize these steps using sufficiently different hardware that cannot be classified as instantiating the same kind of operations. Keeley (2000) thinks this is the case for weakly electric fish; but it does not seem to be the case for IBMs, as their hardware architectures are simply too similar to imply sufficient changes in lower-level types. They share the same sequence of mathematical steps but also the same type of hardware architecture, where tube/transistor difference is not essential to individuating types.

Now, there is yet another, fifth way to think of the computational capacity of computers, namely in terms of how they are connected to all peripheral devices; this will be *information-flow MR*. The pattern of connections is called a "schematic of data flow" in the original IBM documents, and it describes the connections of the central processing unit and magnetic core memory with (1) console lights and switches; (2) cathode ray tube output; (3) magnetic drum storage; (4) and three data syncronizer [sic!] units, each with connections to optional card readers, card punches, printers, tape control units, and other "external signal sources". This way of looking at the computer makes it basically a node in a network of connections, so two realizations would be of the same type of node if and only if they share all the connections. However, if they do, the exact electronic realization of the data flow is basically as irrelevant as in previous cases.

It seems therefore that different perspectives do not really help to see IBMs as multiply realizing the same computational kind. Either we end up with a different computational type for each IBM, as in the instantiation-blueprint MR, or with the same computational and lower-level types for both.

Note that the so-called dimensioned view of realization (Gillett 2002; Wilson and Craver 2007) cannot save the MR claim with regard to this example either, as my argument can be easily rephrased in terms of Cummins-style functional analysis.

The argument is based on the fact that there has to be the same level of grain in individuating both high-level and lower-level types; in other words, we consider lower-level differences to constitute a different type not just in any situation but only when the differences are in a relevant way connected to the way they realize higher-level functionality. One could object that this way, there is no way for MR to occur, but this does not exclude that there are cases of genuine MR. Not at all; the emulator example serves as an instance of genuine MR. It's just that MR occurs only in a very limited number of cases of organizational similarity between computers, so one cannot say that it is in some way essentially linked to the notion of computation the way it was earlier presupposed. In other words, the mechanists should really be skeptical of the role assigned traditionally to MR.

3.3 Organizational Invariance and Substrate Neutrality

At this point, the defenders of MR might still argue that the notion is useful. First, they might insist that there is yet another way of identifying the computational capacities of both IBMs. This is of course possible, and I have definitely not enumerated all possibilities. But I do think that my five options correspond to the usual ways of thinking about computational types. Still, there might be others; I'm not holding my breath to get to know them, however, as it's not so plausible that they might solve the main problem, which is how to have the same level of grain on both levels and still retain MR without begging the question against reductionism.

Another possible objection is to deny that only function-relevant lower-level properties count in identifying the cases of MR. One might bite the bullet and say that the number of threads and the color of the corkscrew do count as realization-relevant properties. The number of cases of MR would definitely increase, as any instance of a type would turn out to be another realization of the type. The question then becomes: Why introduce the notion of realization when we already have the notion of the token? It goes against theoretical parsimony.

Of course, one might point out that I didn't deny that MR is indeed possible and that there are cases of genuine MR. So a reply from a defender of MR might go like this: Maybe it is a bit counterintuitive that these IBM computers are not instances of MR but it's not so vital. We do have genuine MR, so reductionism is doomed anyway. However, I think such an answer would be too quick. I haven't dealt with the question whether MR really warrants antireductionism (and there are authors who stress it does not, cf. (Keeley 2000)), so let me put this question to the side. The intuitive point about MR and computation, one that motivates the claim that MR is somehow essential to computation, is that MR is always logically possible for a given computational system. That may be still true, though achieving MR has a higher price than usually assumed; not only one cannot really make a complex computer out of Swiss cheese or toilet paper, which has been pointed out before (Shagrir 1998), but also one needs to find relevant differences of the proper gain at the lower level. Still, logically this is possible, but no less than creating

a system that exactly simulates the steps in another computational system; there are simply multiple ways one could express similarities and equivalences between computations.

As compared to other precise notions, such as bi-simulation (Malcolm 1996), MR seems to be particularly vague, and there are no facts of the matter that would help decide whether MR occurs or not just because individuation the computational type is inextricably linked to the theoretical interest. This is why the notion needs to be made precise every time by specifying exactly what is meant by the computational type that is supposed to be multiply realized.

I grant that MR might also be defined in the way that such constraints on relevance of realization types and the pragmatic or perspective-bound character of the notion are gone but I do not really see the point of such an endeavor. The problem is that the account of MR should both avoid begging the question against reductionism and not trivialize the notion (by making it effectively as broad as the notion of the token). I think the prospects of such a project are quite gloomy.

So how can one express the intuitive idea that computation is not really linked to the physical substrate in the way many other properties are? IBM 709 and 7090 can indeed compute the same mathematical function and share the same abstract structure of their mechanisms, which is important in explaining and predicting their work. As *purely* computational systems, they are exactly the same, and physical differences are irrelevant, as they make no difference for realization of their capacity to compute the same set of mathematical functions in the same way (by running the same software and being compatible to the same set of peripheral devices, and so forth). For this reason, I suggest that computationalism should rather embrace the notion of *organizational invariance* (or *substrate neutrality*), as both systems share the same *relevant causal topology* on one level of their functioning. Note that under four abovementioned different understanding of computational types (the one related to the instantiation blueprint is an exception) there is organizational invariance between the type and both IBMs.

The notion of organizational invariance has been introduced by David Chalmers (2011) in his theory of physical computation. The notion is defined relative to the causal topology of a system, which is "the abstract causal organization of the system: that is, the pattern of interaction among parts of the system, abstracted away from the make-up of individual parts and from the way the causal connections are implemented." (Chalmers 2011, p. 339) A property is organizationally invariant if it is invariant with respect to the causal topology. In other words, any change to the system that preserves the topology preserves the organizationally invariant property. There is a causal topology of two IBMs that both share, and this topology can be used to define organizationally invariant properties of both. These invariant properties may involve more than causal topology responsible for the computation, as they also have other organizational properties related to the way they are connected to peripheral devices or used in the mainframe lab, for example.

Daniel Dennett's idea of substrate neutrality is similar to organizational invariance and also related to what is intuitive about computation being independent (but not entirely!) from the physical instantiation (Dennett 1995, p. 50). Physically

realized algorithms, according to Dennett, are substrate-neutral in that their causal power is related to their logical structure rather than to particular features of their material instantiation. Now, obviously, without physical instantiation they would not have any causal powers but not all causal powers are relevant for their being computational. In other words, we may safely abstract away from certain physical properties as long as the logical causal powers are still retained, or, to express the same idea in the vocabulary of organizational invariance, as long as the logical causal topology of the computational system is retained.

The computational structure of two IBMs – considered in any of the five ways presented in the Sect. 3.2 – can be retained while we change the physical substrate (in the case of the instantiation blueprint the difference is that IBM 709 does not share the same computational structure as IBM 7090 but there might be, for example, clones of IBM 7090 that do). Of course, not any substrate will be suitable, and because the topology of the whole system has to be retained, one cannot replace one tube with a transistor, or vice versa. But one can create a hybrid IBM 709' that contains a tube-based part and a transistor-based subsystem, as long as one creates a special interface to retain the causal topology of the original IBM. It would still run FORTRAN but with more exotic mixture of hardware.

The idea of substrate neutrality or organizational invariance is that there is a subset of all causal factors in the system which is essential in the system's computational capacity. In the mechanistic framework, the organizationally invariant part is the part responsible for whatever we think is the computational capacity, and one is free to adopt a certain level of abstraction in specifying the whole mechanism (Levy and Bechtel 2013). The paraphrase of the intuitive point about MR will be therefore as follows: The abstract organization of the computational system remains organizationally invariant, or substrate-neutral with respect to a certain level of physical changes. IBM 709 and 7090, considered as physical mechanisms, simply share the same abstract organization, whether it's conceived of as the formal model of computation, the set of computable functions, the sequence of some primitive computable functions, some part of the instantiation blueprint, or the internal and external connections of the computer.

I suspect that many defenders of MR will retort that what I mean by substrate-neutrality is exactly what they mean by MR. That well may be, but this is a conceptual confusion on their part. Substrate-neutrality does *not* imply that there is an additional relevant difference between high- and low-level types that somehow makes cross-level type identification impossible. There is a level of substrate-neutrality for corkscrews, as long material used to build them is sufficiently stiff and so forth. Substrate-neutrality will co-occur with MR in some cases but it has less constraints. Brodman maps do not imply MR but they do imply some substrate-neutrality, as the latter implies that we abstract from some but not all physical detail in describing the relevant causal topology. And we do when we compare anatomical structures for the purposes of topographical mapping.

The acknowledgment of the same abstract organization is not threatened by adopting a reductive stance; irrespective of whether one treats the organization as irreducible or not, it will remain causally relevant, and hence, indispensable in

attaining models of computational mechanisms that are both appropriately general and include necessary specific details. There is simply no reason for refining definitions of MR while we can express the point in a simpler manner, and without begging the question against reductionism. Substrate-neutrality is non-committal, as it does not exclude type-identity of computational and physical kinds. Granted, if one is interested in defeating type-identity and reductionism, then it is not attractive, but it was not my aim to show that computation is not reducible to the physical. To show that it is (nor not) is a task for another paper, and possibly for another author.

Acknowledgements The work on this paper was financed by National Science Centre under the program OPUS, grant no. 2011/03/B/HS1/04563. The author wishes to thank Aaron Sloman for an extended discussion of his idea, to the audience at PT-AT 13, and to the anonymous referee of the previous version of the paper.

References

Aizawa, K., & Gillett, C. (2009). The (multiple) realization of psychological and other properties in the sciences. *Mind & Language, 24*(2), 181–208. doi:10.1111/j.1468-0017.2008.01359.x.

Bechtel, W. (2008). *Mental mechanisms*. New York: Routledge (Taylor & Francis Group).

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science, 66*(2), 175–207.

Block, N. (1990). Can the mind change the world? In G. Boolos (Ed.), *Meaning and method: Essays in honor of Hilary Putnam* (pp. 137–170). Cambridge: Cambridge University Press.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science, 12*, 325–359.

Craver, C. F. (2007). *Explaining the brain. Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, C. F. (2013). Functions and mechanisms: A perspectivalist view. In P. Hunemann (Ed.), *Functions: Selection and mechanisms* (pp. 133–158). Dordrecht: Springer.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72*(20), 741–765.

Davies, P. S. (2001). *Norms of nature: Naturalism and the nature of functions*. Cambridge: MIT Press.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.

Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *The Journal of Philosophy, 65*(20), 627–640.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115. doi:10.1007/BF00485230.

Fresco, N. (2014). *Physical computation and cognitive science*. Berlin/Heidelberg: Springer. doi:10.1007/978-3-642-41375-9.

Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis, 62*(4), 316–323.

Gillett, C. (2011). Multiply realizing scientific properties and their instances. *Philosophical Psychology, 24*(6), 1–12. doi:10.1080/09515089.2011.559625.

Glymour, C. (1994). On the methods of cognitive neuropsychology. *The British Journal for the Philosophy of Science, 45*(3), 815–835. doi:10.1093/bjps/45.3.815.

Haimovici, S. (2013). A problem for the mechanistic account of computation. *Journal of Cognitive Science, 14*(2), 151–181.

IBM Archives: 709 Data Processing System. (2003, January 23). Retrieved January 11, 2014, from http://www-03.ibm.com/ibm/history/exhibits/mainframe/mainframe_PP709.html

IBM Archives: 7090 Data Processing System. (2003, January 23). Retrieved January 11, 2014, from http://www-03.ibm.com/ibm/history/exhibits/mainframe/mainframe_PP7090.html

Keeley, B. L. (2000). Shocking lessons from electric fish: The theory and practice of multiple realization. *Philosophy of Science, 67*(3), 444–465.

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science, 80*(2), 241–261. doi:10.1086/670300.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.

Malcolm, G. (1996). Behavioural equivalence, bisimulation, and minimal realisation. In *Recent trends in data type specification* (pp. 359–378). Berlin/Heidelberg: Springer. doi:10.1007/3-540-61629-2_53.

Meyer, D. E., Osman, A. M., Irwin, D. E., & Yantis, S. (1988). Modern mental chronometry. *Biological Psychology, 26*(1–3), 3–67. doi:10.1016/0301-0511(88)90013-0.

Miłkowski, M. (2011). Beyond formal structure: A mechanistic perspective on computation and implementation. *Journal of Cognitive Science, 12*(4), 359–379.

Miłkowski, M. (2013). *Explaining the computational mind*. Cambridge, MA: MIT Press.

Miłkowski, M. (2014). Computational mechanisms and models of computation. *Philosophia Scientiæ*, *18*(3), 215–228.

Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science, 74*(4), 501–526. doi:10.1086/522851.

Piccinini, G. (2010). Computation in physical systems. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved January 11, 2014 from http://plato.stanford.edu/archives/sum2015/entries/computation-physicalsystems/.

Polger, T. W. (2004). *Natural minds*. Cambridge, MA: MIT Press.

Polger, T. W. (2008). Evaluating the evidence for multiple realization. *Synthese, 167*(3), 457–472. doi:10.1007/s11229-008-9386-7.

Polger, T. W., & Shapiro, L. A. (2008). Understanding the dimensions of realization. *Journal of Philosophy, 105*, 213–222.

Posner, M. I. (2005). Timing the brain: Mental chronometry as a tool in neuroscience. *PLoS Biology, 3*(2), e51. doi:10.1371/journal.pbio.0030051.

Price, C. (2001). *Functions in mind: A theory of intentional content*. Oxford/New York: Clarendon.

Putnam, H. (1975). Philosophy and our mental life. In *Mind, language and reality: Philosophical papers* (Vol. 1, pp. 291–304).

Shagrir, O. (1998). Multiple realization, computation and the taxonomy of psychological states. *Synthese, 114*(3), 445–461. doi:10.1023/A:1005072701509.

Shapiro, L. A. (2000). Multiple realizations. *The Journal of Philosophy, 97*(12), 635–654.

Shapiro, L. A. (2004). *The mind incarnate*. Cambridge, MA: MIT Press.

Shapiro, L. A. (2008). How to test for multiple realization. *Philosophy of Science, 75*(5), 514–525. doi:10.1086/594503.

Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of Science, 66*(4), 542–564.

Wilson, R. A., & Craver, C. F. (2007). Realization: Metaphysical and scientific perspectives. In P. Thagard (Ed.), *Philosophy of psychology and cognitive science* (pp. 81–104). Amsterdam: North Holland. doi:10.1016/B978-044451540-7/50020-7.

Wimsatt, W. C. (2002). Functional organization, analogy, and inference. In A. Ariew, R. Cummins, & M. Perlman (Eds.), *Functions: New essays in the philosophy of psychology and biology* (pp. 173–221). Oxford: Oxford University Press.

Chapter 4
When Thinking Never Comes to a Halt: Using Formal Methods in Making Sure Your AI Gets the Job Done Good Enough

Tarek R. Besold and Robert Robere

Abstract The recognition that human minds/brains are finite systems with limited resources for computation has led researchers in cognitive science to advance the Tractable Cognition thesis: Human cognitive capacities are constrained by computational tractability. As also human-level AI in its attempt to recreate intelligence and capacities inspired by the human mind is dealing with finite systems, transferring this thesis and adapting it accordingly may give rise to insights that can help in progressing towards meeting the classical goal of AI in creating machines equipped with capacities rivaling human intelligence. Therefore, we develop the "Tractable Artificial and General Intelligence Thesis" and corresponding formal models usable for guiding the development of cognitive systems and models by applying notions from parameterized complexity theory and hardness of approximation to a general AI framework. In this chapter we provide an overview of our work, putting special emphasis on connections and correspondences to the heuristics framework as recent development within cognitive science and cognitive psychology.

Keywords Cognitive systems • Complexity theory • Parameterized complexity • Approximation theory • Tractable AI • Approximable AI • Heuristics in AI

T.R. Besold (✉)
The KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, Piazza Domenicani 3, I-39100 Bozen-Bolzano, Italy
e-mail: TarekRichard.Besold@unibz.it

R. Robere
Department of Computer Science, University of Toronto, Toronto, ON, Canada
e-mail: robere@cs.toronto.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_4

4.1 Introduction: The Importance of Formal Analysis for Cognitive Systems Research

After a certain abandonment of the original dream(s) of artificial intelligence (AI) towards the end of the last century, research in cognitive systems, artificial human-level intelligence, complex cognition, and integrated intelligent systems over the last decade has witnessed a revival and is now entering its second spring with several specifically dedicated conference series, symposia, workshops, journals and a growing number of books and high-profile research projects. Still, quite some fundamental questions remain to be answered before a unified approach to solving the big riddles underlying the (re)creation of human-level intelligence and cognition may arise. Currently, there are many different paradigms competing with each other: Symbolism versus connectionism, high-level modeling of specific cognitive capacities versus low-level models with emergent behavior, holistic versus modular approaches.

Each of these paradigms brings along its own terminology, conceptual perspective, and engineering methods, resulting in a wide variety of approaches to solving the intelligence puzzle. This, in turn, makes it hard to establish standards and insights in cognitive models and cognitive systems which are valid on a general level independent of the chosen perspective and methodology. Still, there are a few elements common to most (if not all) of the mentioned approaches (in that, for instance, they are applied in attempts to model one or several human cognitive capacities), making the wish for general principles and results more urgent. Here, formal methods and analyses can provide a solution: Due to their general nature they can often be applied without prior commitment to a particular formalism or architecture, allowing to establish high-level insights and generally applicable findings. In other words, these techniques can provide guidelines and hints at how to unify approaches and progress towards the overall goals of the respective research programs.

In what follows, we give an overview of the status quo of our work on the topic, combining previous independently published contributions and extending the individual pieces into a unified whole. This summary shall provide both evidence supporting the just made claims about the possible role of formal methods for general high-level AI design, and concrete insights concerning heuristics and their use in cognitive systems as important specific example. Section 4.2 introduces the mindset underlying our work before Sect. 4.3 summarizes important theoretical results, followed by a worked application case for our approach in Sect. 4.4. Opening the second half of the chapter, Sect. 4.5 then elaborates the connection between the notion of cognitive heuristics (and their models) to recent results from parameterized complexity and the theory of hardness of approximation, before Sect. 4.6 addresses some of the most common criticisms targeting the application of formal methods to work in AI and cognitive systems. Section 4.7 concludes the chapter, connecting it to related work by other scholars and pointing out some future directions of development.

4.2 Complexity and Cognition

Two famous ideas conceptually lie at the heart of many endeavors in computational cognitive modeling, cognitive systems research and artificial intelligence: The "computer metaphor" of the mind, i.e. the concept of a computational theory of mind as described in Pylyshyn (1980), and the Church-Turing thesis (a familiar version of which is stated in Turing 1969). The former bridges the gap between humans and computers by advocating the claim that the human mind and brain can be seen as an information processing system and that reasoning and thinking correspond to processes that meet the technical definition of computation as formal symbol manipulation, the latter gives an account of the nature and limitations of the computational power of such a system: Every function for which there is an algorithm (i.e., those functions which are "intuitively computable" by a sequence of steps) is computable by a Turing machine—and functions that are not computable by such a machine are to be considered not computable in principle by any machine.

But the "computer metaphor" and the Church-Turing thesis also had significant impact on cognitive science and cognitive psychology. As stated in Cummins (2000), one of the primary aims of cognitive psychology is to explain human cognitive capacities—which are often modeled in terms of computational-level theories of cognitive processes (i.e., as precise characterizations of the hypothesized inputs and outputs of the respective capacities together with the functional mappings between them; cf. Marr (1982) for details). Unfortunately, computational-level theories are often underconstrained by the available empirical data, allowing for several different input-output mappings and corresponding theories. A first attempt at mitigating this problem can now be based on the aforegiven Church-Turing thesis: If the thesis were true, the set of functions computable by a cognitive system would be a subset of the Turing-computable functions. Now, if a computational-level theory would assume that the cognitive system under study computes a function uncomputable by a Turing machine then the model could already be rejected on theoretical grounds.

Still, directly applying the notion of Turing computability as equivalent to the power of cognitive computation has to be considered overly simplistic. Whilst it has been commonly accepted that the thesis holds in general, its practical relevance in cognitive agents and systems is at least questionable. As already recognized in Simon (1957), actual cognitive systems (due to their nature as physical systems) need to perform their tasks in limited time and with a limited amount of space at their disposal. Therefore, the Church-Turing thesis by itself is not strict enough for being used as a constraint on cognitive theories as it does not take into account any of these dimensions. To mitigate this problem, different researchers over the last decades have proposed the use of mathematical complexity theory, like the concept of NP-completeness, as an assisting tool (see, e.g., Levesque 1988; Frixione 2001), bringing forth the so called "*P-Cognition thesis*": Human cognitive capacities are hypothesized to be of the polynomial-time computable type.

However, using the "polynomial-time computable" as synonymous with "efficient" may already be overly restrictive. In modern times there are many examples of problems which have algorithms that have worst-case exponential behaviour, but tend to work quite well in practice on small inputs (take, for example, any of the modern algorithms for the travelling salesperson problem). However, this is not the type of restriction that we will focus on. Instead we take the following viewpoint: it is often the case that we as humans are able to solve problems which may be hard in general but suddenly become feasible if certain parameters of the problem are restricted. This idea has been formalized in the field of *parameterized complexity theory*, in which "tractability" is captured by the class of fixed-parameter tractable problems FPT[1]:

Definition 4.1 (FPT). A problem P is in FPT if P admits an $O(f(\kappa)n^c)$ algorithm, where n is the input size, κ is a parameter of the input constrained to be "small", c is an independent constant, and f is some computable function.

Originating from this line of thought, van Rooij (2008) introduces a specific version of the claim that cognition and cognitive capacities are constrained by the fact that humans basically are finite systems with only limited resources for computation: Applying the just presented definition from parameterized complexity theory, this basic notion of resource-bounded computation for cognition is formalized in terms of the so called "FPT-*Cognition thesis*", demanding for human cognitive capacities to be fixed-parameter tractable for one or more input parameters that are small in practice (i.e., stating that the computational-level theories have to be in FPT).

4.3 Theoretical Foundation: The Tractable AGI Thesis

But whilst the aforementioned P-Cognition thesis also found its way into AI (cf., e.g., Cooper 1990; Nebel 1996), the FPT-Cognition thesis this far has widely been ignored. Recognizing this as a serious deficit, for example in Robere and Besold (2012) and Besold and Robere (2013a), we proposed a way of (re)introducing the idea of tractable computability for cognition into AI and cognitive systems research by rephrasing and accordingly adapting the FPT-form of the Tractable Cognition thesis. As all of the currently available computing systems used for implementing cognitive models and cognitive systems are ultimately finite systems with limited resources (and thus in this respect are not different from other cognitive agents and human minds and/or brains), in close analogy we developed the "Tractable AGI thesis" (Tractable Artificial and General Intelligence thesis).

[1]For an introduction to parameterized complexity theory see, e.g., Flum and Grohe (2006) and Downey and Fellows (1999).

Tractable AGI thesis Models of cognitive capacities in artificial intelligence and computational cognitive systems have to be fixed-parameter tractable for one or more input parameters that are small in practice (i.e., have to be in FPT).

Concerning the interpretation of this thesis, suppose a cognitive modeler or AI system designer is able to prove that his model at hand is—although in its most general form possibly NP-hard—fixed-parameter tractable for some set of parameters κ. This implies that if the parameters in κ are fixed small constants for problem instances realized in practice, then it is possible to efficiently compute a solution.

4.4 Worked Example: Complex Analogies in HDTP

Before further continuing the theoretical line of work in Sect. 4.5, we want to spend some time on a worked application case of analyzing a cognitive AI system by means of formal methods. We therefore give a fairly detailed reproduction of results from a parameterized complexity study of the Heuristic-Driven Theory Projection (HDTP) computational analogy-making framework (originally presented in Robere and Besold 2012). By this we hope to show how the mostly academic-theoretical considerations from Sects. 4.2 and 4.3 directly connect to everyday AI and cognitive systems practice.

4.4.1 The Motivation Behind It

During the course of a day, we use different kinds of reasoning processes: We solve puzzles, play instruments, or discuss problems. Often we will find ourselves in places and times in which we apply our knowledge of a familiar situation to the (structurally similar) novel one. Today it is undoubted that one of the basic elements of human cognition is the ability to see two a priori distinct domains as similar based on their shared relational structure (i.e., analogy-making). Some prominent cognitive scientists as, for example, Hofstadter (2001), go as far as to consider analogy the core of cognition itself. Key abilities within everyday life, such as communication, social interaction, tool use, and the handling of previously unseen situations crucially rely on the use of analogy-based strategies and procedures. One of the key mechanisms underlying analogy-making, relational matching, is also the basis of perception, language, learning, memory and thinking, i.e., the constituent elements of most conceptions of cognition (Schwering et al. 2009b).

Because of this crucial role of analogy in human cognition, researchers in cognitive science and artificial intelligence have been creating computational models of

analogy-making since the advent of computer systems. But the field has changed significantly during that time: Early work such as that of Reitman et al. (1964) or Evans (1964) should serve as a proof of concept for the possibilities and the power of AI systems, possibly paving the way for more flexible approaches to reasoning and artificial cognition. Still, from a theoretical and methodological point of view these systems were not necessarily committed to considerations concerning cognitive adequacy or psychological plausibility and did not correspond to a fully developed underlying theoretical paradigm about human analogy-making. In contrast, modern analogy systems—the most prominent of which probably is the Structure-Mapping Engine (SME, Falkenhainer et al. 1989) and MAC/FAC (Gentner and Forbus 1991)—come with their respective theory about how analogy-making works on a human scale (see, e.g., Gentner (1983) for the theory behind SME) and often even make claims not only on a computational level of description, but even hypothesize more or less precisely specified algorithmic mechanisms of analogy.

And this now is where our proposed approach for formal analysis comes into play: If human-likeness in a system's theoretical foundations and behavior is assumed, it should also automatically become clear that the same standards of evaluation and the same formal properties which are true for human cognition have to be met and have to hold for the system. Moreover this has to be the case in general and not only on a selected subset of examples or under positively limiting conditions and a priori assumptions on the possible cases a system might encounter (unless, of course, these assumptions can also be made without loss of generality for the human counterpart). Analogy-making is a prime example for this setting as—precisely due to the aforementioned variety of occurrences and manifestations of this cognitive capacity—the architect of a cognitive system model of analogy has to make sure that certain properties of the system hold true with a high degree of independence from the specific problem case at hand.

Furthermore, from a tractability perspective, computational analogy systems are a first-rate application area for our theoretical paradigm. Analogy-making has gained attention in cognitive science and cognitive AI not only because of its general applicability but also due to the fact that humans seem to be able to retrieve and use analogies in very efficient ways: Conversations happen in real time, social interaction—although highly diverse in its different levels—is pervasive and in most cases does not require attention or conscious thought (even if we are only acquainted with the general rules and paradigm and not with the specific situations we encounter), and once we understood the solution to a certain riddle or problem we have no problem immediately applying it to analogical cases even when they are completely different in appearance or setting.

4.4.2 The Formal Analysis

Heuristic-Driven Theory Projection, introduced in Schwering et al. (2009a), is a formal theory and corresponding software implementation, conceived as a mathematically sound framework for analogy-making. HDTP has been created for

computing analogical relations and inferences for domains which are given in the form of a many-sorted first-order logic representation. Source and target of the analogy-making process are defined in terms of axiomatizations, i.e., given by a finite set of formulae. HDTP tries to produce a generalization of both domains by aligning pairs of formulae from the two domains by means of a process called *anti-unification*, which tries to solve the problem of generalizing terms in a meaningful way, yielding for each term an "anti-instance" in which some subterms have been replaced by variables (which in turn would allow for a retrieval of the original terms by a substitution of the variables by appropriate subterms).

HDTP in its present version uses a restricted form of higher-order anti-unification presented in Krumnack et al. (2007). In higher-order anti-unification, classical first-order terms are extended by the introduction of variables which may take arguments (where classical first-order variables correspond to variables with arity 0), making a term either a first-order or a higher-order term. Then, anti-unification can be applied analogously to the original first-order case, yielding a generalization subsuming the specific terms. The class of substitutions which are applicable in HDTP is restricted to (compositions of) the following four cases: renamings (replacing a variable by another variable of the same argument structure), fixations (replacing a variable by a function symbol of the same argument structure), argument insertions, and permutations (an operation rearranging the arguments of a term).

This formalism has proven capable of detecting structural commonalities not accessible to first-order anti-unification, as for instance also structural commonalities between functions and predicates within the logical language can be found and exploited (whilst the first-order formalism would in these be limited to the respective argument positions only), allowing for a more general recognition of relational mappings (as opposed to mere attribute mappings). Once the generalization has been computed, the alignments of formulae together with the respective generalizations can be read as proposals of analogical relations between source and target domain, and can be used for guiding an analogy-based process of transferring knowledge between both domains (see Fig. 4.1 for an overview of the analogy-making process). Analogical transfer results in structure enrichment on the target side, which corresponds to the addition of new axioms to the target theory, but may also involve the addition of new first-order symbols.

Whilst HDTP undoubtedly exhibits pleasant properties—say, in terms of expressivity of the modeling language, and clarity of the underlying conceptual approach—until recently there had been no detailed analysis of its computational

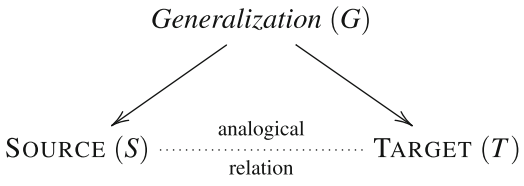


Fig. 4.1 A schematic overview of HDTP's generalization-based approach to analogy

tractability. In order to change this unsatisfactory state of affairs, we decided to apply some techniques from parameterized complexity theory to the system trying to better understand its strengths and weaknesses.

As already mentioned, a restricted higher-order anti-unification is defined as any composition of a certain set of unit substitutions which can formally be specified in the following way:

Definition 4.2 (Restricted higher-order anti-unification). The following are the types of unit substitutions allowed in restricted higher-order anti-unification.

1. A renaming $\rho(F, F')$ replaces a variable $F \in \mathcal{V}_n$ with another variable $F' \in \mathcal{V}_n$:

$$F(t_1, \ldots, t_n) \xrightarrow{\rho(F,F')} F'(t_1, \ldots, t_n).$$

2. A fixation $\phi(F, f)$ replaces a variable $F \in \mathcal{V}_n$ with a function symbol $f \in \mathcal{C}_n$:

$$F(t_1, \ldots, t_n) \xrightarrow{\phi(F,f)} f(t_1, \ldots, t_n).$$

3. An argument insertion $\iota(F, F', V, i)$ is defined as follows, where $F \in \mathcal{V}_n, F' \in \mathcal{V}_{n-k+1}, V \in \mathcal{V}_k, i \in [n]$:

$$F(t_1, \ldots, t_n) \xrightarrow{\iota(F,F',V,i)} F'(t_1, \ldots, t_{i-1}, V(t_i, \ldots, t_{i+k}), t_{i+k+1}, \ldots, t_n).$$

   It "wraps" k of the subterms in a term using a k-ary variable, or can be used to insert a 0-ary variable.
4. A permutation $\pi(F, \tau)$ rearranges the arguments of a term, with $F \in \mathcal{V}_n$, $\tau : [n] \to [n]$ a bijection:

$$F(t_1, \ldots, t_n) \xrightarrow{\pi(F,\tau)} F(t_{\pi(1)}, \ldots, t_{\pi(n)}).$$

A *restricted substitution* is a substitution which results from the composition of any sequence of unit substitutions.

By considering different combinations of restricted substitutions we can define several different forms of higher-order anti-unification. Unfortunately, as already recognized by Krumnack et al. (2007), the least general generalizer is not necessarily unique. Therefore, in our analysis we instead consider decision versions of the problems parameterized by the number of substitutions, variables, and types of variables used.

Problem 4.1 (F Anti-Unification).

Input: Two terms f, g, and a natural $k \in \mathbb{N}$
Problem: Is there an anti-unifier h, containing at least k variables, using only renamings and fixations?

Problem 4.2 (FP Anti-Unification).

Input: Two terms f, g, and naturals $l, m, p \in \mathbb{N}$.
Problem: Is there an anti-unifier h, containing at least l 0-ary variables and at least m higher arity variables, and two substitutions σ, τ using only renamings, fixations, and at most p permutations such that $h \overset{\sigma}{\rightarrow} f$ and $h \overset{\tau}{\rightarrow} g$?

Problem 4.3 (FPA Anti-Unification).

Input: Two terms f, g and naturals $l, m, p, a \in \mathbb{N}$.
Problem: Is there an anti-unifier h, containing at least l 0-ary variables, at least m higher arity variables, and two substitutions σ, τ using renamings, fixations, at most p permutations, and at most a argument insertions such that $h \overset{\sigma}{\rightarrow} f$ and $h \overset{\tau}{\rightarrow} g$?

We summarize our (parameterized) complexity-theoretic results of higher-order anti-unification in the following theorem[2]:

Theorem 4.1. *1. F Anti-Unification is solvable in polynomial time.*
2. *FP Anti-Unification is* NP*-complete and* W[1]*-hard w.r.t. parameter set* $\{m, p\}$.
3. *Let r be the maximum arity and s be the maximum number of subterms of the input terms. Then FP Anti-Unification is in* FPT *w.r.t. parameter set* $\{s, r, p\}$.
4. *FPA Anti-Unification is* NP*-complete and* W[1]*-hard w.r.t. parameter set* $\{m, p, a\}$.

4.4.3 Interpretation of the Results

We want to provide some thoughts on the consequences of the complexity results from the previous section, putting the obtained insights into a cognitive AI context and thereby making the intrinsic connection between the formal considerations and the analogy mechanism of the implemented system explicit.

We focus on a result directly affecting HDTP. The result showing that FP higher-order anti-unification is W[1]-hard gives a hint at the difficulty introduced by the operations admissible within the restricted higher-order anti-unification on the complexity of the analogy-making process. Indeed, the only way that FP anti-unification can restructure the order of the terms is by argument permutations,

[2]The corresponding proofs of the respective results can be found in Robere and Besold (2012). Moreover, in the theorem statements W[1] refers to the class of problems solvable by constant depth combinatorial circuits with at most 1 gate with unbounded fan-in on any path from an input gate to an output gate. In parameterized complexity, the assumption W[1] $\neq$ FPT can be seen as analogous to P $\neq$ NP.

and our results show that even allowing a *single* permutation is enough to imply computational hardness. If we contrast this result against the polynomial-time algorithm for F anti-unification, we have evidence that even a slight ability to restructure the input terms makes higher-order anti-unification a difficult problem to solve.

Now, additionally making the (most likely reasonable) assumption that $\mathsf{P} \neq \mathsf{NP}$ (and $\mathsf{FPT} \neq \mathsf{W}[1]$) holds, the presented hardness results cast a shadow on the suitability of the HDTP framework in its present state as basis for a general model for high-level cognitive capacities or a general cognitive architecture. Still, it should be noticed that the mere fact that HDTP in its present state is basically intractable does not mean that future versions cannot be made tractable. Here, the insights obtained from the formal analysis can serve as guidelines for the future evolution of the system: In our opinion, one of the main questions for future theoretical research in relation to HDTP will have to address the question of how HDTP's version of computing generalizations via restricted higher-order anti-unification can be further constrained in a meaningful way as to obtain maximal expressivity and applicability whilst still staying within the domain of polynomial solvability. Also, more parametrized analysis will be needed, showing which are the factors that really impact complexity, and which are aspects of a problem that are not really harmful.

4.5 Setting Limits to Heuristics in Cognitive Systems

Leaving the basic considerations and the application study of formal means of analysis of cognitive AI systems behind us we want to return to a more theoretical part of our work. A still growing number of researchers in cognitive science and cognitive psychology, starting in the 1970s with the "heuristics and biases" program (Kahneman et al. 1982), and today prominently heralded, for instance, in the work of Gigerenzer and colleagues (2011), argues that humans in their common sense reasoning do not apply any full-fledged form of logical or probabilistic reasoning to possibly highly complex problems, but instead rely on mechanisms—which are mostly automatic and unconcious—that allow them to circumvent the impending complexity explosion and nonetheless reach acceptable solutions to the original problems. Amongst the plethora of proclaimed automatisms are, for example, the representativeness heuristic (Kahneman et al. 1982) or the take-the-best heuristic (Czerlinski et al. 1999).

All of these mechanisms are commonly subsumed under the all-encompassing general term "heuristics". Still, on theoretical grounds, at least two quite different general types of approach have to be distinguished within this category: Either the complexity of solving a problem can be reduced by reducing the problem instance under consideration to a simpler (but solution equivalent) one, or the problem instance stays untouched but—instead of being perfectly (i.e., precisely) solved—is dealt with in a good enough (i.e., approximate) way.

Now, taking the perspective of an architect of a cognitive system considering to include human-inspired heuristics in his reasoning model for solving certain tasks, a crucial question quite straightforwardly arises: Which problems can actually be solved by applying heuristics—and how can the notion of heuristics be theoretically modeled on a sufficiently high level as to allow for a general description? Having a look at recent work in parameterized complexity theory and in hardness of approximation, we find that the two distinct types of heuristics naturally correspond to two well-known concepts from the respective fields. As firstly shown in Besold (2013) and reproduced in the following, this opens the way for establishing a solid theoretical basis for models of heuristics in cognitive systems.

4.5.1 The Reduction Perspective

In Sect. 4.3, the Tractable AGI thesis demanded for models of cognitive capacities in AI to be in FPT. However, there is also a non-trivial corollary that can be derived from this property: any instance of a problem in FPT can be reduced to a *problem kernel*.

Definition 4.3 (Kernelization). Let P be a parameterized problem. A kernelization of P is an algorithm which takes an instance x of P with parameter κ and maps it in polynomial time to an instance y such that $x \in P$ if and only if $y \in P$, and the size of y is bounded by $f(\kappa)$ (f a computable function).

Theorem 4.2 (Kernelizability Downey et al. 1997). *A problem P is in* FPT *if and only if it is kernelizable.*

This theorem on the one hand entails that any positive FPT result obtainable for the model in question essentially implies that there is a "downward reduction" for the underlying problem to some sort of smaller or less-complex instance of the same problem, which can then be solved—whilst on the other hand (assuming $\mathsf{W}[1] \neq \mathsf{FPT}$) any negative result implies that there is no such downward reduction. This equivalence forms a first connecting point to some of the different heuristics frameworks in cognitive science and cognitive psychology—and can also have important ramifications for the modeling of many cognitive capacities in computational cognitive systems.

On the one hand, from a constructive perspective, by actively considering the kernelizability of problems (or rather problem classes), we can provide inspiration and first hints at hypothesizing a specialized cognitive structure capable of computing the reduced instance of a problem, which then might allow for an efficient solving procedure. On the other hand, taking a more theoretical stance, by categorizing problems according to kernelizability (or their lack thereof) we also can establish a distinction between problem classes which are solvable by

reduction-based heuristics and those which are not—and can thus already a priori decide whether a system implementing a reduction-based heuristics might generally be unable to solve a certain problem class.

From a theoretical perspective the strict equivalence between FPT-membership and kernelizability of a problem is somewhat surprising. However, on practical and applied grounds, the correspondence should seem natural and, moreover, should fairly directly explicate the connection to the notion of reduction-based heuristics: If cognitive heuristics are as fast and frugal as commonly claimed, considering them anything but (at worst) polynomial-time bounded processes seems questionable. But now, if the reduced problem shall be solvable under resource-critical conditions, using the line of argument from Sects. 4.2 and 4.3, we can just hope for it to be in FPT. Now, combining the FPT-membership of the reduced problem with the polynomial-time complexity of the heuristics, already the original problem had to be fixed-parameter tractable. Still, reduction-based heuristics are not trivialized by this: Although original and reduced problem are in FPT, the respective size of the parameters may still differ between instances (which possibly can make an important difference in application scenarios for implemented cognitive systems).

4.5.2 The Approximation Perspective

There is also a complementary perspective offering an alternate possibility of (re)interpreting heuristics, namely the theory of approximation algorithms: Instead of precisely solving a kernel as proposed by reduction-based heuristics, compute an approximate solution to the original problem (i.e., the solution to a relaxed problem). The idea is not any more to perfectly solve the problem (or an equivalent instance of the same class), but to instead solve the problem to some "satisfactory degree".

Here, a candidate lending itself for being considered a standard analogous to FPT in the Tractable AGI thesis is APX, the class of problems allowing polynomial-time approximation algorithms:

Definition 4.4 (APX)**.** An optimization problem P is in APX if P admits a constant-factor approximation algorithm, i.e., there is a constant factor $\epsilon > 0$ and an algorithm which takes an instance of P of size n and, in time polynomial in n, produces a solution that is within a factor $1 + \epsilon$ of being optimal (or $1 - \epsilon$ for maximization problems).

Clearly, here the meaningfulness and usefulness of the theoretical notion in practice crucially depends on the choice of the bounding constant for the approximation ratio: If the former is meaningfully chosen with respect to the problem at hand, constant-factor approximation allows for quantifying the "good enough" aspect of the problem solution and, thus, offers a straightforward way of modeling the notion of "satisficing" (Simon 1956) (which in turn is central to many heuristics considered in cognitive science and psychology).

One should also be careful to note that "constant-factor approximation" is also quite unrestrictive in ways other than pure tractability. While this class captures problems that may have efficient algorithms which produce an solution that is, say, half as good as an optimal solution, it also contains problems that have efficient $1/1000$-approximations, or $1/1{,}000{,}000$. While these approximation factors almost never appear in practice, they are theoretically allowed. However, we believe in principal that this serves to strengthen any *negative* results which would place problems outside of APX.

As in the case of the reduction-based heuristics, one of the main advantages of formally identifying approximation-based heuristics with APX lies in its limiting power: If a problem shall be solved at least within a certain range from the optimal solution, but it turns out that the problem does not admit constant-factor approximation for the corresponding approximation parameter, the problem can a priori be discarded as unsolvable with approximation-based heuristics (unless one wants to also admit exponential-time mechanisms, which might be useful in selected cases but for simple complexity considerations does not seem to be feasible as a general approach).[3]

4.5.3 Joining Perspectives

Having introduced the distinction between reduction-based and approximation-based heuristics, together with proposals for a formal model of the mechanisms behind the respective class, we now want to return to a more high-level view and look at heuristics in their entirety. This is also meaningful from the perspective of the initially mentioned system architect: Instead of deciding whether he wants to solve a certain type of task applying one of the two types of heuristics and then conducting the corresponding analysis, he might just want to directly check whether the problem at hand might be solvable by any of the two paradigms. Luckily, a fairly recently introduced theoretical concept allows for the integration of the two different views—FPT and APX can both be combined via the concept of fixed-parameter approximability and the corresponding problem class FPA:

Definition 4.5 (FPA)**.** The fixed-parameter version P of a minimization problem is in FPA if—for a recursive function f, a constant k, and some fixed recursive function g—there exists an algorithm such that for any given problem instance I

[3] On the other hand, considering more restrictive notions than APX as, for instance, PTAS (the class of problems for which there exists a polynomial-time approximation scheme, i.e., an algorithm which takes an instance of a optimization problem and a parameter $\epsilon > 0$ and, in polynomial time, solves the problem within a factor $1 + \epsilon$ of the optimal solution) does not seem meaningful to us either, as also human satisficing does not approximate optimal solutions up to an arbitrary degree but in experiments normally yields rather clearly defined cut-off points at a certain approximation level.

with parameter k, and question $OPT(I) \leq k$, the algorithm which runs in $O(f(k)n^c)$ (where $n = |I|$) either outputs "no" or produces a solution of cost at most $g(k)$.

As shown in Cai and Huang (2006), both polynomial-time approximability and fixed-parameter tractability with witness (see Cai and Chen (1997) for details) independently imply the more general fixed-parameter approximability. And also on interpretation level FPA artlessly combines both views of heuristics, at a time in its approximability character accommodating for the notion of satisficing and in its fixed-parameter character accounting for the possibility of complexity reduction by kernelizing whilst keeping key parameters of the problem fixed.

Clearly, the notion of fixed-parameter approximability is significantly weaker than either FPT and kernelization, or APX. Nonetheless, its two main advantages are the all-encompassing generality (independent of the type of heuristics) and yet again the introduction of a categorization over problem types: If problems of a certain kind are not in FPA, this also excludes membership in any of the two stricter classes—and thus (in accordance with the lines of argument given above) in consequence hinders solvability by either type of heuristics.

Recalling the Tractable AGI thesis introduced in Sect. 4.3, we can use the just outlined conception of FPA for not only considering classical strict processing and reasoning in cognitive systems, but for also accounting for models of cognitive heuristics. This allows us to adapt the original thesis into a (significantly weaker but possibly more "cognitively adequate") second form:

Fixed-Parameter Approximable AGI thesis Models of cognitive capacities in artificial intelligence and computational cognitive systems have to be fixed-parameter approximable for one or more input parameters that are small in practice (i.e., have to be in FPA).

Whilst the original Tractable AGI thesis was aimed at AI in general (thus also including forms of high-level AI which may not be human-inspired, or which in their results shall not appear human-like), the just postulated thesis, due to its strong rooting in the (re)implementation of human-style processing, explicitly targets researchers in cognitive systems and cognitive AI.

4.6 The Importance of Formal Analysis for Cognitive Systems Research Revisited

An often heard fundamental criticism of trying to apply methods from complexity theory and formal computational analysis to cognitive systems and cognitive models are variations of the claim that there is no reason to characterize human behavior in

terms of some "problem" (i.e., in terms of a well-defined class of computational tasks), from this drawing the conclusion that computational complexity is just irrelevant for the respective topics and fields.[4] In most cases, this judgement seems to be based on either a misconception of what a computational-level theory in the sense of Marr (1982) is (which we will refer to as "*description error*"), or a misunderstanding of what kind of claim complexity theory makes on this type of theory (in the following referred to as "*interpretation error*").

The *description error* basically questions the possibility of describing human behavior in terms of classes of computational tasks. The corresponding argument is mostly based on the perceived enormous differences between distinct manifestations of one and the same cognitive capacity already within a single subject (at the moment for descriptive simplicity's sake—without loss of generality—leaving aside the seemingly even more hopeless case of several subjects). Still, precisely here Marr's Tri-Level Hypothesis (i.e., the idea that information processing systems should be analyzed and understood at three different—though interlinked—levels) comes into play. Marr proposed three levels of description for a (biological) cognitive system, namely a computational, an algorithmic, and an implementation level. Whilst the latter two are concerned with how a system does what it does from a procedural and representational perspective (algorithmic level), and how a system is physically realized (implementation level), the computational level takes the most abstract point of view in asking for a description of what the system does in terms of giving a function mapping certain inputs on corresponding outputs. So what is needed to specify a computational-level theory of a cognitive capacity[5] is just a specified set of inputs, a set of corresponding outputs (each output corresponding to at least one element from the set of inputs) and a function establishing the connection between both (i.e., mapping each input onto an output). But now, due to the high degree of abstraction of the descriptive level, this allows us to characterize human cognitive capacities in general in terms of a computational-level theory by specifying the aforementioned three elements—where inputs and outputs are normally provided (and thus defined) by generalization from the real world environment, and the function has to be hypothesized by the respective researcher.[6]

[4]For reasons unclear to the authors this perspective seems to be more widespread and far deeper rooted in AI and cognitive systems research than in (theoretical) cognitive science and cognitive modeling where complexity analysis and formal computational analysis in general by now have gained a solid foothold.

[5]Here we presuppose that cognitive capacities can be seen as information processing systems. Still, this seems to be a fairly unproblematic claim, as it simply aligns cognitive processes with computations processing incoming information (e.g., from sensory input) and resulting in a certain output (e.g., a certain behavioral or mental reaction) dependent on the input.

[6]Fortunately, this way of conceptualizing a cognitive capacity naturally links to research in artificial cognitive systems. When trying to build a system modeling one or several selected cognitive capacities, we consider a general set of inputs (namely all scenarios in which a manifestation of the cognitive capacity can occur) which we necessarily formally characterize—although maybe only implicitly—in order to make the input parsable for the system, hypothesize a function mapping inputs onto outputs (namely the computations we have the system apply to the inputs) and finally

Once all three parts have been defined, formal computational analyses can directly be conducted on the obtained computational-level theory as the latter happens to coincide in form with the type of problem (i.e., formal definition of a class of computational tasks) studied in computational complexity and approximation theory. And also the existence of at least one computational-level theory for each cognitive capacity is guaranteed: Simply take the sets of possible inputs and corresponding outputs, and define the mapping function element-wise on pairs of elements from the input and output, basically creating a lookup table returning for each possible input the respective output.

Leaving the description error behind us, we want to have a look at the *interpretation error* as further common misconception. Even when modifying the initial criticism by not questioning the overall possibility of characterizing human behavior in terms of classes of computational tasks, but rather by stating that even if there were these classes, it would not have to be the case that humans have to be able to solve all instances of a problem within a particular class, we believe that this argument is missing the point. First and foremost, as further elaborated upon in the initial paragraph of the following section, we propose to use complexity and (in)-approximability results rather as a safeguard and guideline than as an absolute exclusion criterion: As long as a computational-level theory underlying a computational cognitive model is in FPT, APX, or FPA—where in each case the system architect has to decide which standard(s) to use—the modeler can be sure that his model will do well in terms of performance for whatever instance of the problem it will encounter.[7] Furthermore, it is clear that in cognitive systems and cognitive models in general a worst-case complexity or approximability analysis for a certain problem class only rarely (if at all) can be taken as an absolute disqualifier for the corresponding computational-level theory.[8] It might well be the case that the majority of problem instances within the respective class is found to be well behaving and easily solvable, whilst the number of worst-case instances is very limited (and thus possibly unlikely to be encountered on a basis frequent enough as

obtain a well-characterized set of outputs (namely all the outputs our system can produce given its programming and the set of inputs).

[7]In discussions with researchers working in AI and cognitive systems very occasionally critical feedback relating to the choice of FPT, APX, and FPA as reference classes has been given, as these have (curiously enough) been perceived as too less restrictive. Harshly contrasting with the previously discussed criticism it was argued that human-level cognitive processing should be of linear complexity or less. Still, we do not see a problem here: Neither are we fundamentalist about this precise choice of upper boundaries, nor do we claim that these are the only meaningfully applicable ones. Nonetheless, we decided for them because they can quite straightforwardly be justified and are backed up by close correspondences with other relevant notions from theoretical and practical studies in cognitive science and AI.

[8]Of course this also explicitly includes the case in which the considered classes are conceptually not restricted to the rather coarse-grained hierarchy used in "traditional" complexity theory, but if also the significantly finer and more subtle possibilities of class definition and differentiation introduced by parametrized complexity theory and other recent developments are taken into account.

to turn their occurrence into a problem). However, at a high level, complexity theory can still provide researchers with meaningful information—given a computational intractability or inapproximability result the researcher has an opportunity to refocus his energies onto algorithms or analysis which are more likely to be fruitful. Moreover, in the process of the formal analysis, the researcher now becomes more intimately familiar with the problem at hand—which parameters of the problem are responsible for a "complexity explosion", which parameters can be allowed to grow in an unbounded fashion and still maintain computational efficiency. And, of course, if one is still put off by this sort of complexity analysis, it may simply be a matter of changing the particular analysis type: Where worst-case analyses may on certain grounds be questionable as decisive criterion about the overall usefulness of a particular computational-level theory for a cognitive capacity, average-case analyses (which admittedly are significantly harder to perform) can change the picture dramatically.

4.7 Conclusion: Limiting the Limits

A second frequent criticism (besides the popular general objection discussed in the previous section) against the type of work presented in this paper is that demanding for cognitive systems and models to work within certain complexity limits might always be overly restrictive: Maybe each and every human mental activity actually is performed as an exponential-time procedure, but this is never noticed as the exponent for some reason always stays very small. Undoubtedly, using what we just presented, we cannot exclude this possibility—but this also is not our current aim. What we want to say is different: We do not claim that cognitive processes are without exception within FPT, APX, or FPA, but we maintain that as long as cognitive systems and models stay within these boundaries they can safely be assumed to be plausible candidates for application in a resource-bounded general-purpose cognitive agent (guaranteeing a high degree of generalizability, scalability, and reliability). Thus, if a system architect has good reasons for plausibly assuming that a particular type of problem in all relevant cases only appears with a small exponent for a certain exponential-time solving algorithm, it may be reasonable to just use this particular algorithm in the system. But if the architect should wonder whether a problem class is likely to be solvable by a resource-bounded human-style system in general, or if it should better be addressed using reduction-based or approximation-based heuristics, then we highly recommend to consider the lines of argument presented in the previous sections.

Concerning related work, besides the conceptually and methodologically closely related, but in its focus different efforts by van Rooij (2008) and colleagues in theoretical cognitive science (see, e.g., Kwisthout and van Rooij 2012; Blokpoel et al. 2011), of course there also is work relating fixed-parameter complexity to AI. Still, except for very few examples as, e.g., Wareham et al. (2011), the applications mostly are limited to more technical or theoretical subfields of artificial

intelligence (see, e.g., Gottlob and Szeider (2008) for a partial survey) and—to the best of our knowledge—this far have not been converted into a more general guiding programmatic framework for research into human-level AI and cognitive systems. To a certain extent, a laudable exception to this observation may be found in Chapman (1987), where the author presents a high-level algorithm for general purpose planning and—using formal methods similar to the ones considered above—derives general constraints for domain-independent planning under certain assumptions on the expressivity of the action representations, together with ways of avoiding the found limitations.

We therefore in our future work hope to develop the overall framework further, also showing the usefulness and applicability of the proposed methods in different worked examples from several relevant fields: The range of eligible application scenarios spans from models of epistemic reasoning and interaction, over cognitive systems in general problem-solving scenarios, down to models for particular cognitive capacities as, for example, analogy-making (see, e.g., Sect. 4.4 and additionally Besold and Robere (2013b) for a proof of concept).

References

Besold, T. R. (2013). Formal limits to heuristics in cognitive systems. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems (ACS) 2013*, Baltimore.

Besold, T. R., & Robere, R. (2013a). A note on tractability and artificial intelligence. In K. U. Kühnberger, S. Rudolph, & P. Wang (Eds.), *Artificial General Intelligence – 6th International Conference, AGI 2013, Proceedings*, Beijing (Lecture Notes in Computer Science, Vol. 7999, pp. 170–173). Springer.

Besold, T. R., & Robere, R. (2013b). When almost is not even close: remarks on the approximability of HDTP. In K. U. Kühnberger, S. Rudolph, & P. Wang (Eds.), *Artificial General Intelligence – 6th International Conference, AGI 2013, Proceedings*, Beijing (Lecture Notes in Computer Science, Vol. 7999, pp. 11–20). Springer.

Blokpoel, M., Kwisthout, J., Wareham, T., Haselager, P., Toni, I., & van Rooij, I. (2011). The computational costs of recipient design and intention recognition in communication. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, Boston (pp. 465–470)

Cai, L., & Chen, J. (1997). On fixed-parameter tractability and approximability of {NP} optimization problems. *Journal of Computer and System Sciences, 54*(3), 465–474. doi:http://dx.doi.org/10.1006/jcss.1997.1490.

Cai, L., & Huang, X. (2006). Fixed-parameter approximation: Conceptual framework and approximability results. In H. Bodlaender & M. Langston (Eds.), *Parameterized and exact computation* (Lecture Notes in Computer Science, Vol. 4169, pp. 96–108). Berlin/Heidelberg: Springer. doi:10.1007/11847250_9.

Chapman, D. (1987). Planning for conjunctive goals. *Artificial Intelligence, 32*(3), 333–377.

Cooper, G. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence, 42*, 393–405.

Cummins, R. (2000). "How does it work?" vs. "What are the laws?" two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117–145). Cambridge: MIT.

Czerlinski, J., Goldstein, D., & Gigerenzer, G. (1999). How good are simple heuristics? In G. Gigerenzer, P. Todd, & the ABC Group (Eds.), *Simple heuristics that make us smart*. New York: Oxford University Press.

Downey, R. G., & Fellows, M. R. (1999). *Parameterized complexity*. New York: Springer.

Downey, R. G., Fellows, M. R., & Stege, U. (1997). Parameterized complexity: A framework for systematically confronting computational intractability. In *Contemporary Trends in Discrete Mathematics: From DIMACS and DIMATIA to the Future*. Providence: AMS.

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. In *Proceedings of the April 21–23, 1964, Spring Joint Computer Conference AFIPS '64 (Spring)* (pp. 327–338). New York: ACM. doi:http://doi.acm.org/10.1145/1464122.1464156

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*(1), 1–63. doi:10.1016/0004-3702(89)90077-5.

Flum, J., & Grohe, M. (2006). *Parameterized complexity theory*. Berlin: Springer.

Frixione, M. (2001). Tractable competence. *Minds and Machines, 11*, 379–397.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Gentner, D., & Forbus, K. (1991). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science, 19*, 141–205.

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundation of adaptive behavior*. New York: Oxford University Press.

Gottlob, G., & Szeider, S. (2008). Fixed-parameter algorithms for artificial intelligence, constraint satisfaction and database problems. *The Computer Journal, 51*(3), 303–325. doi:10.1093/comjnl/bxm056.

Hofstadter, D. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 499–538). Cambridge: MIT.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge/New York: Cambridge University Press.

Krumnack, U., Schwering, A., Gust, H., & Kühnberger, K. (2007). Restricted higher-order anti-unification for analogy making. In *Twentieth Australian Joint Conference on Artificial Intelligence*. Berlin: Springer.

Kwisthout, J., & van Rooij, I. (2012). Bridging the gap between theory and practice of approximate bayesian inference. In *Proceedings of the 11th International Conference on Cognitive Modeling*, Berlin (pp. 199–204).

Levesque, H. (1988). Logic and the complexity of reasoning. *Journal of Philosophical Logic, 17*, 355–389.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing visual information*. San Francisco: Freeman.

Nebel, B. (1996). Artificial intelligence: A computational perspective. In G. Brewka (Ed.), *Principles of knowledge representation* (pp. 237–266). Stanford: CSLI Publications.

Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundation of cognitive science. *The Behavioral and Brain Sciences, 3*, 111–132.

Reitman, W. R., Grove, R. B., & Shoup, R. G. (1964). Argus: An information-processing model of thinking. *Behavioral Science, 9*(3), 270–281. doi:10.1002/bs.3830090312.

Robere, R., & Besold, T. R. (2012). Complex analogies: Remarks on the complexity of HDTP. In *Twentyfifth Australasian Joint Conference on Artificial Intelligence* (Lecture Notes in Computer Science, Vol. 7691, pp. 530–542). Berlin/New York: Springer.

Schwering, A., Krumnack, U., Kühnberger, K. U., & Gust, H. (2009a). Syntactic principles of heuristic-driven theory projection. *Journal of Cognitive Systems Research, 10*(3), 251–269.

Schwering, A., Kühnberger, K. U., & Kokinov, B. (2009b). Analogies: Integrating multiple cognitive abilities – guest editorial. *Journal of Cognitive Systems Research 10*(3), 175–177.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129–138. doi:10.1037/h0042769.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: Wiley.

Turing, A. (1969). Intelligent machinery. In B. Meltzer & D. Michie (Eds.), *Machine intelligence* (Vol. 5, pp. 3–23). Edinburgh: Edinburgh University Press.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science, 32*, 939–984.

Wareham, T., Kwisthout, J., Haselager, P., & van Rooij, I. (2011). Ignorance is bliss: A complexity perspective on adapting reactive architectures. In *Proceedings of the First IEEE Conference on Development and Learning and on Epigenetic Robotics*, Frankfurt am Main (pp. 465–470).

Chapter 5
Machine Intelligence and the Ethical Grammar of Computability

David Leslie

Abstract Since the publication of Alan Turing's famous papers on "machine intelligence" over six decades ago, questions about whether complex mechanical systems can partake in intelligent cognitive processes have largely been answered under the analytical rubric of their capacity successfully to simulate symbol-mongering human behavior. While this focus on the mimetic potential of computers in response to the question "Can machines think?" has come to be accepted as one of the great bequests of Turing's reflections on the nature of artificial intelligence, I argue in this paper that a closer look at Turing's oeuvre reveals an especially informative tension between the pragmatic and normative insights, which enabled him in 1936 to formulate his pioneering version of the theory of mechanical computability, and his later attempt to argue for a simplistic notion of "machine intelligence" as an effectual imitation of the human mind. In fleshing out the source of this tension, I endeavor to show how the mimetic model of "thinking machines" that Turing eventually embraces is ultimately at cross-purposes with the normative-pragmatic insights by which he reached his original innovations in computability theory and combinatorial logic.

Keywords Machine intelligence • Effective calculability • Church-turing thesis • Turing test • Computability • Inexhaustibility • Limitative theorems • Metamathematics

5.1 Introduction

It is an astonishing feature of the glut of articles, books and collected volumes published over the last three decades on the foundations of the notion of computability that so few have placed the so-called "confluence of ideas of 1936" (Gandy 1988), which yielded the Church-Turing Thesis, in the wider intellectual-historical context of the converging philosophical revolutions that were erupting contemporaneously

D. Leslie (✉)
Harvard University, 59 Shepard Street, Cambridge, MA 02138, USA
e-mail: davidleslie@fas.harvard.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_5

both in European and in Anglo-American thinking. Beyond the formalist challenge to finitary proof spurred by Hilbert's program, beyond the nettle and tug of the thorn of diagonalization, which had kept unresolved the disturbing antinomies seemingly inherent in the new logic, "what was in the air"[1] in decades surrounding '36 – an air inhaled in common by Post, Church, Turing and Gödel but also by Peirce, James, Dewey, the later Wittgenstein, Husserl and Heidegger – was the waning mist of a Cartesian bygone ever more displaced by the gathering density of a wholly unprecedented postmetaphysical milieu. In this weighty atmosphere, fundamental questions then circulating about the epistemic and ontological status of formal systems, about the nature of meaning and the anatomy of predication, and about the proper character of the relationship of mind and world were being refracted through the deflationary prism of incipient pragmatist, hermeneutic, and phenomenological insights.

Common to all of these latter perspectives was a radical rejection of the last remnants of the paradigm of representationalism that had all but defined the explanatory aspirations of the heritage of thinkers who wrote in the wake of Descartes' inauguration of the modern "mind's" search for certain knowledge – a search to find the indubitable building blocks of human comprehension whose mentalistic starting point shunted all interrogatory intent through the methodological sluice gate of the pregiven dichotomy of *intellectus et rei*. This heritage had become all-too-enthralled by what C.I. Lewis (1929) was to call the *ignis fatuus* (p. 9), the alluring but illusory glow, of an absolute Reality to which Cartesian thoughts or Kantian categories or Tractarian facts could be said to correspond. That world "out there" against which incorrigible minds, conceptual schemes, and picture-theoretic syntax could be set in adequational relief gradually became a "world well lost" (Rorty 1982) insomuch as the plodding self-realization of the limitations of a reason increasingly situated in its own practical, cultural and historical contexts of origination, seemed only more and more to confirm that there was no "relation of closeness of fit" between the semiotic systems and the patterns of signs and relations through which humans thought and communicated and some timeless "way the world is" (Goodman 1960). And, as Occidental reason thus came progressively to discover its own predicative and ontological limitations so too did it come progressively to debunk the "spectator theory of knowledge" (Dewey 1960, p. 23) and the reifying myths of mental- and world-givenness, which had enabled the Cartesian tradition under all of its motley empiricist and rationalist guises to satiate its own epistemological urge.

Two critical trajectories, in particular, burst forth from within modern thinking itself and conspired to instigate the radical deflation of such a metaphysically bloated Cartesian ambition. In the first place, there arose an intensifying proceduralization of reason viewed not as a hypostatized *res cogitans*, *viz.* not as thinking-stuff or mental essence, but rather as a concrete mode of a particular kind

[1]This is Gandy's phrase and the theme around which he organizes "The Confluence of Ideas in 1936."

of shared human practice.[2] Fueled by the exemplary model of empirical scientific method first conceptualized in the work of thinkers like Gassendi, Locke and Newton, this sociocultural shift to the detranscendentalized paradigm of procedural rationality gradually deflated the explanatory hopes of any theoretical outlook which had thenceforth been steered by a sense of privileged metaphysical purport. Instead, the thoroughgoing but salutary skepticism, which had initially fueled the rise of the fallibilistic self-understanding of the "experimental" and inductively-driven modern sciences, seemed increasingly to penetrate all Occidental modes of inquiry and explanation. A splinter of insuperable provisionalism thus broke off in the skin of post-traditional thinking. As a consequence, the ontological linking point of "mind and world," which had been required by Cartesian metaphysics, dissolved into the dialogically animated linking point of reciprocally implicating processes of giving and asking for reasons. Indeed, it was based upon this procedural transposition of reason on the plane of the everyday communicative exchange of evidence, assessment and conclusion that the collaborative quest for the clarification of common and publically available experience began to define a postconventional society compelled to reproduce itself without recourse to any non-self ascribed social or ontological categories. The strengthening acknowledgement of this processual thawing of reason hence signaled a dramatic shift in intellectual orientation whereupon the collaborative coping of co-operating humans, who were responsible solely to each other in creating and passing on the shared vocabularies and conventions, which steered and coordinated their practices, began to take elucidatory priority over the simplistic copy-theoretic views of a fixed and ineffable Reality that was somehow intrinsically liable to being captured by the polished mirror of the "Western mind." In this respect, the antecedent normative significance of reasoning as a consensually anchored social practice came radically to dislodge the residua of naïve essentialist and imagist logic, which, by theoretical fiat, had erroneously reified the provinces of Thought and Being.

In the second place and congruently, a modern reasoning hence ever more compelled to ground itself *ex nihilo* steadily came to discover the difficult and self-defeating irony emergent from its own insuperable finitude. That is, in the wake of the deflationary impetus most forcefully initiated by Kant's anthropocentric turn, the embodied carrier of modern knowing became the condition of possibility of knowledge as such, hence mooring the sovereign claims of human reasoning to the fraught transience of its all-too-human medium. Here, the preeminent subject of knowledge, that very ascendant source of sapient knowing, that sole station of the Renaissance "dignity of man" and of the Cartesian *cogito*, was itself called before the critical tribunal of its own corporeally and sociohistorically situated enabling conditions. And, as the sinews of experience thus violently turned back upon the flesh of experience itself, reason's claim to the unconditionality force of its assertions was improbably situated in an animate cross-reference with itself. That is, it was situated

[2]This "theme of postmetaphysical thinking" has been stressed both by Jürgen Habermas (1992) and by Karl-Otto Apel (1998).

in all the conditioning contexture of its own transitory circumstances of genesis. In this sense, aside from yielding the strains of historicism and detranscendentalization that culminated in the epoch-defining materialist critique of metaphysics, which sent tremors through the intellectual landscape of *fin-de-sicle* Europe, Kantian critical philosophy incited the difficult and painfully reflexive learning process that Michel Foucault (1966/1994) terms the "finitization of man," "the primary discovery of finitude" (p. 342) *per se* – what Hans Georg Gadamer (1960/1992), in a similar vein, calls a *pathei mathos*, an agonizing but formative education in epistemic humility and in the inexorability of mortal limitation (p. 357).

And so, it is in this context of the radical finitization and procedural deflation of reason in the midst of a fading trace of Cartesian epistemological convictions that I would like, in this paper, to reposition the innovations of the protagonists of '36. For it is their insurgent skepticism towards Hilbert's militant rejection of the *ignoramibus*, their efforts to rescale the ambitions of metamathematics and logic by revealing "absolutely unsolvable combinatory problems," that stimulated the progressive fine-graining of the idea of effective calculability which culminated in Turings famous paper on computable numbers. To be sure, inasmuch as the development of the Church-Turing Thesis was motivated by the pressure to devise negative and balloon-bursting responses to Hilbert's program, the very concepts of general recursiveness, of lambda-definability, of finite combinatorial processes, and of Turing machines all functioned as subsidiary but requisite components of the more basic critical-deflationary endeavor to yield with unadulterated clarity the incompleteness and undecidability results for which the thinkers of '36 are rightfully best remembered. And, it is far from coincidental, as I will claim here, that what made Turing computability so definitive, indeed conclusive, for Turing's contemporaries was the distinctive and wholly novel way that he was able to apply concepts of radical finitization and procedural deflation to this most basic kind of mathematical reasoning.

In keeping with this wider conceptual-historical purview, I want to focus, in what follows, on an especially informative tension that surfaces in the writings of Turing – a tension between the pragmatic and normative insights, which enabled him in the 1930s to formulate both his pioneering version of the theory of mechanical computability as well as his solution to Hilbert's *Entscheidungsproblem*, and his later attempt to argue for a simplistic notion of "machine intelligence" as an effectual imitation of the human mind, an attempt founded on a recidivistic appeal to a mimetic logic largely parasitic upon the exact Cartesian heritage against which his earlier detransendentalizing insights in no small measure militated.

I will, in effect, attempt to tell here a tale of two Turings. The Turing of '36, on my account, revolutionizes recursion theory by finitizing effective calculability as a universalizable process of agentially-situated and intersubjectively compositional reckoning, which is both constitutively subject to the detranscendentalizing constraints of the concrete practice of human computing and ultimately limited by the radical indeterminacy and incompleteness immanent in mathematical experience *as such*. By contrast, the Turing of the 1950s, while making several significant qualifications which intimate the limitations of machine intelligence, still presses a reifying notion of human mindedness into the service of a one-dimensional copy

theory of AI. In fleshing out the source of the tension between these two Turing's, I will endeavor to show how the mimetic model of "thinking machines" that Turing eventually embraces is ultimately at cross-purposes with the normative-pragmatic insights by which he reached his original innovations and for reason of which he later waivers with regard to the unrestricted scope of precisely that imitative model. I will attempt finally to make explicit just what was so implicitly and intuitively convincing to his contemporaries about this turn to generalizable finite procedures of computing as a way better to articulate the nature of recursive calculation than any offered before.

5.2 The Lures of Imitation

It is now over six decades since Alan Turing first answered the question of whether complex mechanical systems can partake in intelligent cognitive processes under the analytical rubric of their capacity successfully to simulate symbol-mongering human behavior. Much of the downstream controversy that has surrounded Turing's famous test involves the criterion of adequacy that he presupposes as establishing what counts as valid evidence for machine intelligence. Does equating the phrase "programming a machine to think" with the phrase "programming a machine *to imitate* a brain" (as Turing does in his 1951 lecture, "Can Digital Computers Think?" (1951/2004, p. 485)) constitute a convincing inferential step in arguing for AI? More precisely, is the *analysans* of this logic of *mimesis* or imitation adequate to the *analysandum* of predicating of machines that they possess intelligence? Can the capacity for thinking and understanding be warrantably projected onto computers thereby?

One of the immediate challenges faced by those who would answer any one of these questions affirmatively is the problem of latent hypostatization evidently forced upon any supporter of the "simulation-view." The issue is relatively straightforward: the mimetic criterion of machine intelligence must assume that there exists some-*thing* denoted "intelligence," "mind," or "thought" that can be imitated as such. That is, the logic of mimesis must assume the concreteness of the object of simulation and is therefore apt to reify an animate social process as a mechanically duplicable product without ever justifying why or how it is entitled to do so. Such a potentially incautious conflation of the "thinking" and the "thought" makes this position susceptible to what Max Black (1946) called a "process/product ambiguity" (p. 177) and Sellars later (1956/1997), the "notorious ing/ed ambiguity" (p. 54).[3] Along the lines of these latter views, the imitation-theoretic use of the concept of intelligence relies on the vagueness created by its own inattention to the difference between the activity or force of intellecting and the results or artifacts of what is intellected, in order to deem intelligence traits that are speciously concrete

[3]Notably, a distinction of this sort arises much earlier in analyses of logic but then again even more forcefully in the process philosophy of Alfred North Whitehead. It takes on a more semantic complexion in the work of Black, Sellars, and Searle (1968, p. 422).

and thereby subject to replication without remainder. This suppositional basis is consequently vulnerable to the charge that thinking is a human activity that is done amidst living and interacting people who must together continuously rebuild the ship of collaboratively achieved meaning in the open seas of indeterminate human experience. It is not a *concretum in rerum natura* that is found and that is thereby liable to simple-minded copying. In this regard, the tender underbelly of the imitation argument first formulated by Turing inheres in exactly the way that such a Cartesian recidivism enables a kind of essentialist epistemic overstretching whereby "the mental" once again comes to be erroneously predicable as such and therefore erroneously available for totalizing simulation.

Be that as it may, it is quite remarkable and not-often-enough noted that throughout his elaborations of the nature of machine intelligence, Turing himself is explicitly at grips with the problems inherent in this mimetic way of thinking and offers several significant qualifications of his position. The most vital of these for our purposes arises in Turing's 1948 discussion in "Intelligent Machinery" of the limitations of discipline and the necessity for initiative in the generation of intelligent behavior. In an oft-quoted passage of this essay, he writes,

> To convert a brain or machine into a universal machine is the extremist form of discipline. Without something of this kind one cannot set up proper communication. But discipline is certainly not enough to produce intelligence. That which is required in addition we call initiative. This statement will have to serve as a definition. Our task is to discover the nature of this residue as it occurs in man and to try and copy it in machines (1948/2004, p. 429).

Now, I want to reformulate the line of inquiry Turing is opening here as follows: What are the conditions of possibility of the animation of intelligent behavior? Beyond the structuring mechanism of algorithmic pattern, which functions merely to program intelligence, what are the dynamics of mobilization which actually quicken intellectual power as such?

Interestingly enough, Turing (1948) begins to broach this puzzle by substituting the more nuanced term "search" for that of "initiative." The shift enables him a higher degree of analytical precision in distinguishing between three distinct sources of animation which he calls "forms of search" and taxonomizes as "evolutionary," "intellectual" and "cultural" (p. 430). Leaving aside his very brief Darwinian allusion to the first of these, let me instead focus on explicating a crucial point of contact between the latter two, for it is precisely this nexus between these intellectual and cultural residua that cuts into the copy-theoretic view of machine intelligence perhaps in a deeper and more profound way than Turing himself seems to have realized.

While Turing coins the phrase "intellectual search" in the 1948 paper in order to indicate the exploratory impetus of a human mindedness forever compelled to pursue novel combinatory relationships beyond the strictures of mechanical discipline, he already indicates in his 1938 dissertation on ordinal logics that the source of the unremitting pressure behind this pursuit is the discovery of Gödel. By diagonalizing out of the enumerable set of all computable functions, Gödel is able to show, for Turing, "the impossibility of finding a formal logic which wholly eliminates the necessity of using intuition" (1938/2004, p. 193). Put differently,

in light of the quandary of incompleteness, the ultimate burden of mathematical reasoning is perforce placed on the shoulders of the practicing and intuiting mathematician, who is obliged to set about an open-ended intellectual quest insofar as he aspires to reason mathematically at all. The modest beginning of any and all mathematical inquiry is, in this sense, the splinter of living doubt irremissibly lodged in the dermis of a reason congenitally inept to fulfill its charge.

From the start, then, Turing derives the animating factor underlying "intellectual search" from a certain negative characterization which is anchored in the predicament of inexhaustibility first exposed under the aegis of Gödel's arithmetization of metamathematics (Gödel 1931/1965); intellectual search, that is, is made necessary by the particular manner in which the problems of unsolvability and undecidability proved by the limitative theorems underwrite what Gödel (1951) refers to as "the incompletability of mathematics" (1951/1995, p. 133). The irreducible residue of initiative as this indeterminate intellectual quest precluded from recourse to any sort of *calculus ratiocinator* is thus animated in virtue of the very finitizing bounds to algorithmic reasoning apprehended by the ingenuity and intuitive judgment of human thinking itself.

But herein lies a manifestation of the difficult irony of finitization I mentioned at the start. A mathematical reasoning turned back upon itself in the throes of self-referential perplexity discovers the truth of its own limitations. However, the determination of this limitation is already, in two significant senses, a transgression of that same limitation: Either the transgression signifies the ascendency of a human mind capable of proving the limits of its own rule-bound modality of reasoning simultaneously from within and outside of the boundaries of reason itself, or, the transgression signifies something of the radical openness and indeterminacy of a mathematical experience incapable of resting satisfied with the completion of its task – a mathematical reasoning hence humbled by the fact that not one of its embodied carriers can have the last word, leaving each of them no choice but to speak, to ask each other "why?", to demand further reasons.

Turing, it seems to me, sides with the latter position. The limitative results do not, on his view, reveal the superiority of the human mind vis-à-vis the mechanical discipline of machines (as the mathematical objection to AI à la Gödel appeared to him in 1950 to maintain (1950/2004, p. 450)). Rather they signal the necessity of the residue of initiative, which in his last published work on solvable and unsolvable problems, Turing (1954/2004) equates with a "common sense" that is irreducible to the exercise of reason (1954/2004, p. 595).

Turing's use of the term "common sense" here takes on a rich and complex significance equal to the viscosity of its own etymological provenance. For, inasmuch as the ongoing and incompletable task of mathematics ranges well-beyond the efforts of any single person, the task must itself be conceived as a common and collaborative effort. It is the effort of a kind of unbounded community of search. The "support of common sense" to which Turing refers hence enlists the kind of intersubjective texture already present in Aristotle when he writes of the exceptional human quality of *synaisthanomenoi* (common sense, sensing-in-common). As Aristotle argues, unlike cows or sheep, which merely dwell upon

their pastures side-by-side, human beings sense-together as a primitive mode communicative sharing.[4] That is, they coexist as con-sensing and as undergoing the inexhaustible and sociohistorically-anchored trial to con-sense ever more. In this respect, the *sensus communis* – that common sense to which Horace, Vico, and Lord Shaftesbury would each in their own way later refer as an intesubjectively self-attuning reservoir of collective understanding – reverberates into Turing's own appropriation of the phrase.

And, it is in this denser meaning of "common sense" that the point of contact I have suggested to obtain between intellectual and cultural searches inheres. That is to say, these two animating sources of the motility of intelligence coalesce in the inexhaustible residue of initiative instantiated in an unbounded community of search. In his explication of the cultural form of search, Turing (1948) expresses this convergence best,

> ...the isolated man does not develop any intellectual power. It is necessary for him to be immersed in an environment of other men, who's techniques he absorbs for the first 20 years of his life. He may then perhaps do a little research of his own and make a very few discoveries which are passed on to other men. From this point of view the search for new techniques must be regarded as carried out by the human community as a whole, rather than by individuals (1948/2004, p. 431)

This intricated picture of the socioculturally-embedded sources of the "residue of initiative" enables us, at present, to think with Turing against Turing. For, if the conditions of possibility of intelligent behavior involve a finitizing compulsion to limitless search animated exceptionally from within the human community as an obliged mode of sharing, if such a species-ramifying quest is hence instantiated in what Karl Jaspers (1953) terms the "challenge to unbounded communication" (p. 19), it becomes well-nigh impossible to conceive just how such a residue of initiative, such a living context of relevance, can simply be copied in machines. This, we might call the animation problem of AI: Insofar as the necessary preconditions of the animation of intelligence entail an ongoing and forever-provisional public-process of redeeming reasons amidst living interlocutors compelled collectively to cope with the predicament of their finitude in light of the thorn of inexhaustibility simultaneously broken in their discursive skin, intelligence itself, must, in a strict sense, be an ethical-practical and normative concept. It must be a concept animated singularly amidst speaking and interacting human beings, who, taken together, are compelled to assume the conjoint burden of intersubjectively clarifying experience, and who, taken individually, are compelled to assume responsibility for the unique, albeit transient, contributions each of them may make thereunto. Intelligence involves, as John Haugeland (1998) memorably puts it, an "existential commitment" to the boundless project of understanding (pp. 340ff.), an interminable pledge to take responsibility in the face of the other for the sake of a greater conspecific whole.

[4]This occurs in *Nicomachean Ethics* 1170a28–1171b35. See also (Agamben 2009, pp. 25–38).

5.3 The Ethical Grammar of Computability

Now, I want to suggest, at this point, that in excavating these implicit moral-practical underpinnings upon which Turing's mature notions of "initiative" and "search" are secured, we are led to something of a remarkable parallelism in Turing's thinking writ large. That is, the ethical-pragmatic preconditions of the animation of intelligence I have been unearthing basically rearticulate some of the core insights which subtend Turing's earlier theory of computability and which, as I will argue shortly, helped to make the latter so convincing and revolutionary.

As is well known, Turing computability erupts onto the mathematical scene of recursion theory and combinatorial logic as a way out of the *cul-de-sac* evidently faced by the thinkers of '36, who were attempting to figure out how to justify the identification of effective calculability with recursiveness. A more precise and non-circular picture of this connection promised a clearer definition of the notion of formal systems as well as of the micro-mechanics of stepwise proof governed by theorem predicates. For the group of theorist in the trenches of this problem, the gamut of available logico-deductive methods seemed to have been exhausted, for, given the pool of meta-theoretical concepts accessible to them at the time, there appeared to be no convincing way to warrant the placement of recursiveness restrictions on the inferential steps of effective calculations.

The appearance of Turing's "On Computable Numbers," hence heralded nothing less than a sea change in this fundamental quarter of logic and mathematics, for Turing introduced a radically pragmaticizing reorientation of the question of computability itself. Such a reorientation brought the investigation of effective calculability closer to its own native phenomenon, namely, that of concrete combinatory processes as they occur in the activity of human calculating; it therefore expanded the scope and efficacy of the explication of the nature of computing at the same time as it deflated the epistemological stretch of the terms of its explanation. As Robin Gandy (1988) has pointed out, Turing reaches his innovations in defining effective computability by transposing the ontological question "What is a computable function?" onto the practical plane of the adverbial question "What are the possible processes which can be carried out in computing a real number?" (p. 80). This pragmatic turn from the knowing-what to the knowing how of calculability consequently prompts a twofold finitization of the very notion of computation.

First, computational intelligibility is now conceived as factoring down to procedures of understanding which are concretely operative in embodied thinking processes. It is characterized, that is, strictly by the practical stepwise application of a finite number of basic recursive relations or rules to strings of recognizable symbols. Effective methods of computing are graspable, in this way, solely by means of the reflexively available mental resources common to all human computors and are limited by the particular locality and boundedness constraints to which the latter are subject in virtue of their determinate physical capabilities.

Secondly, in keeping with this latter aspect of the agentially-situated generalizability of computation, effective calculability can now be classified as intersubjectively compositional. Computations are henceforth to be marked out by their basic capacity to secure "radical intersubjectivity" (Sieg 2006, p. 205). Stephen Kleene (1987; 1988) and Douglas Hofstadter (1979/1994) accordingly place this second intersubjective implication of the pragmatic turn inaugurated by Turing under the rubric of what they term a "public process version of the Church-Turing Thesis."[5] According to the latter, the method of effective calculability is bound by the proviso that it can be communicated reliably from one sentient being to another by means of language, in a finite amount of time, and with complete case-to-case iterability (Hofstadter 1979/1994, p. 557).

Taken together, these aspects of the agentially-situated generalizability and the intersubjective compositionality, which underlie all processes of effective calculation, seem to suggest that Turing's epochal 1936 insights into mechanical procedures are, most elementally, insights into the normative character of the social practice of computing. To be sure, the agentially-situated generalizability of calculation implicitly secures the authority of each human computor; that is, it enables individual computors to take responsibility for the assertability of their claims and to count before one another equally as being entitled to do so. Correspondingly, the intersubjective compositionality of computation implicitly secures the accountability of computors to each other. It safeguards that each is held responsible and is compelled to settle assertoric accounts through the reciprocal assessment of the soundness and coherence of his or her inferential judgments.[6] In view of this primordial entanglement of authority and responsibility at the site of the social animation of computing, agentially-situated generalizability and intersubjective compositionality can be seen as ethical-pragmatic presuppositions of computation as a mode of con-sensing, as a way of semiological sharing. Along these lines as well, the Church-Turing Thesis can be viewed, in its most basic aspect, as a rational reconstruction of the ethical-practical preconditions of the communicatively-instantiated inter-activity of computation.[7]

[5] Specifically in Kleene (1987, pp. 493–494), Kleene (1988, p. 50), and Hofstadter (1979/1994, 556).

[6] The elemental role of these dimensions of authority and responsibility in the practice of giving and asking for reasons has been stressed most recently in the writings of Robert Brandom. On his view, the undertaking of inferentially articulated commitments in dialogical processes of meaning redemption constitutes the basis of the normative-pragmatics of rational communication. In the stress he places upon this normative character of the sociality of reason, Brandom writes in a broadly Kantian-Hegelian heritage and, in so doing, joins the ranks of Humboldt, Peirce, Sellars, Apel, and Habermas. See especially his (1994), (2000), (2009, pp. 52–77).

[7] There is, in fact, a third step we must take here to flesh out fully the ethical-pragmatic preconditions of the social practice of computing, but one to which, in keeping with the scope of this paper, I can only allude. I want to suggest that, by situating the central limitative claims set forth by Turing (1936/2004) within the context of the reconstruction of these preconditions I have been offering, we can begin to discern a largely unrecognized bridging concept that links the programmatic significance of the Church-Turing Thesis to that of the undecidability results

This latter acknowledgement of what we might call the ethical grammar of computability, allows us, I want to argue, to shed a new hue of light on the long contested issue of the epistemological status of the Church-Turing Thesis. It is undeniable that Turing's claims instantly revolutionized the self-understanding of logic and metamathematics for the thinkers of '36. Church himself maintained at the time that Turing computability, "has the advantage of making the identification with effectiveness evident immediately" (1937, p. 43) And, Gödel, of course, over the next three decades would continually sing the praises of Turing's formulation as a "precise and unquestionably adequate definition of the general notion of formal system,"[8] "an absolute definition of an interesting epistemological notion, i.e. one not depending on the formalism chosen" (1946/1965, p. 84). Indeed, for Gödel, Turing computability allows both the metamathematical absoluteness claim of Gödel Reckonability and Theorems VI and XI of the incompleteness papers to be thoroughly clarified and given a completely general expression.

But the issue of just what made Turing's thesis so immediately evident, unquestionably adequate, and hence universally convincing has remained unclear. Notwithstanding their recognition of the profound expressive ramifications of Turing's insights, Gödel and Church, for instance, both leave the answer to this question in almost total obscurity. The so-called "human version" of Church-Turing inspired by the work of Gandy and fleshed out later by Wilfried Sieg addresses this issue by viewing the adequacy of the position as emerging from its codification or idealization of the limited capacities of human calculators. As Sieg (2002) argues, in examining human mechanical computability, Turing, "exploited limitations of the

Turing achieved in offering his negative answer to the *Entscheidungsproblem*. As is well-known, Turing accomplished the latter by, as he put it, correctly applying "the diagonal process" (Turing 1936/2004, p. 72) to what has since come to be known as the halting problem. Once Turing had established a perspicuous definition of an algorithm under the rubric of his machines, it then became possible for him, first, to Gödelize an enumerable set of the latter by arithmetizing them and, then, to apply the tool of diagonalization in order to exploit the operative limitations exposed by the fate of self-referential foundering thereby met in the *process of computing*. It is in this pragmaticizing "application of the diagonal process" that, I want to suggest, we can identify a third precondition of the social practice of computing. By resituating the predicament of unsolvability in the tangible milieu of the human process of calculating, Turing shows that the embodied computor is always subject to a certain irreducible factor of alterity (instantiated in the iterative constituents of the anti-diagonal sequence). The latter both determines the absolute limitation of the finite calculative practice and transgresses that very limitation in virtue of the unincorporability of the antidiagonal sequence into the diagonal process which delimits that calculative practice as such. The factically manifesting architectonic of undecidability, which Turing derives therefrom, can be seen as an enabling condition of the radical openness of mathematical experience, a condition of possibility of mathematical possibility, if you will. It animates the unbounded exigency to mathematical communication and operates, in turn, as a precondition of the social practice of calculating more primitive than intersubjective compositionality and agentially-situated generalizability inasmuch as these latter two are themselves spurred by the barb of living doubt emergent from such a predicament of unsolvability per se.

[8]Gödel appended this comment in 1963 to his 1931, "On Formally Undecidable Propositions of the Principia Mathematica and Related Systems" (Gödel 1986, p. 191).

human computing agent to motivate restrictive conditions" (p. 395). This empirical-conditioning view, however, falls short of completely achieving the aim of its own explanatory aspiration, for, inasmuch as it merely re-describes, further elaborates and axiomatizes how Turing reached his result, it fails to make fully explicit why it is that that set of empirical restrictions enables us properly to link embodied computability with the normativity of the proof conditions themselves. That is, if, as I want to maintain, the restrictive conditions of the proof predicates that bind effective calculation to recursiveness derive from the normative requirements effective as ethical-pragmatic preconditions of the social practice of computing, then "Sieg constraints," as we might call them, are simply a codification of the concrete phenomenological milieu wherein those practices subsist and not a justification of the normative mechanism operating therein.

From an opposite starting point, Saul Kripke (2013) has recently proposed a reduction of the Church-Turing Thesis to a "special form of mathematical argument," specifically, to Hilbertian conditions of stepwise deduction in first-order logic with identity (p. 80). But, in this reformalizing move, Kripke takes a not-so-virtuous step back into circularity, for, in attempting to return the Church-Turing Thesis to the stage of the Gödel completeness of first-order logic and thereby to recast effective computability as a pure form of deducibility, he merely resets the "stumbling block" (Sieg 1994, p. 78) Turing had so innovatively kicked away. By logicians edict, as it were, he posits a domain of mathematical computation cut away from all-too-human strictures and communicative contexts. And, from the standpoint of this Kripkean view from nowhere, computability theory is seemingly consigned to teeter between a Scylla of phenomenological vacuity: "I wish to exclude questions of empirically or physically based computation… Here we are talking about mathematical computation, whether done by a human being, or a machine, or anything else" (2013, p. 89). And a Charybdis of unsupportable solipsism: "It does not seem to me to be particularly relevant that the directions [guiding an effective computation] be public, that more than one person is involved. It suffices that someone give a finite set of directions to herself" (2013, p. 89).

What the almost symmetrically opposing shortfalls of Sieg's finiteness conditions and the Kripkean reformalization of "first-order algorithms" show us, I want to suggest, is a certain explicatory deficit common to both perspectives – one, which leads us back to the ethical grammar version of the Church-Turing Thesis, now as a kind of reduction theorem. That is to say, the determinacy, boundedness and locality conditions distilled in Sieg's conceptual analysis set restrictive conditions on the process of effective computability, which are significant on the plane of justifying the Church-Turing Thesis only insofar as they function to underwrite, at the descriptive and empirical level, a reconstruction of the ethical-practical preconditions of computability as a mode of communicating. In the same way, Kripke's formalist appeal to the frictionless mechanisms of combinatory rules and stepwise deducibility can play an equivalent justifying role only to the extent that they bring light to the normative strictures discharged epistemologically in the discursive processes of justification and assessment for which mathematical interlocutors are argumentatively liable and by which they are able to bind others to

the force of the reasons they offer. Both Sieg's and Kripke's positions are therefore in their own respective ways indubitably correct but, on my account, only partially so.

5.4 Conclusion

Let me conclude here by suggesting that, beyond Turing's innovation, an argument for the ethical grammar of computability can be derived immanently from within the extant dialogue surrounding the remarkable confluence of ideas of 1936, for adumbrations of such a grammar had been implicit in these diverse but converging articulations from their very beginnings. Hilbert's program, which played such a central role in motivating these latter developments, arose as a challenge for finitistic illumination amidst the darkness of the long shadow cast by the antinomies upon which the nineteenth century movement toward the arithmetization of analysis had seemingly run aground. However, it was a challenge that counterintuitively sought out reflexive resources for the enlightenment to which it aspired exclusively from within the presupposed procedural parameters set by that same movement. And herein lies the salient point: The development of arithmetization first spurred by the critical impetus to rigor gathering in the work of Gauss, Abel and Cauchy and brought to fruition in the results of Weierstrass, Cantor and Dedekind, heralded a radical shift away from any argumentative recourse to hazy geometrical or empirical intuitions.[9] *Per contra*, it demanded a move toward the precision of intersubjectively vindicable eidetic constructions composed entirely of natural numbers and finite and infinite systems of them. The arithmetization movement hence called for a shift away from the sorts of characterizations of mathematical objects and relations that were prone to subjectivist deception and empirical vagary toward the justificatory sovereignty of integers. This, in turn, prompted the shift to a focus on the consistency of models of axiomatically delimited notions and formal systems, which were liable to communicative transmission through pellucid semiological means. The explosion of axiomization, which would consequently follow, also triggered growing ambitions to reveal a consistent and purely logical foundation of the number system – ambitions, in effect, to apply the finitistically-derived and publically-available transparency of the combinatorial iterability of integers in order to ground the totality of integers itself. The formal constraints on inferential practice set by this movement towards arithmetization "all the way down" hence had an implicit normative character, *ab initio*, one set in motion by a certain subjacent ethical-critical pressure to ever more unencumbered and rationally-clarified ways of eidetic sharing.

Of course, under its logicist and Platonist guises, this totalizing drive to arithmetization eventually foundered, not in virtue of its critical-finitist impulses

[9]For good overviews of this movement towards arithmetization, see Murawski (1999) and Kleene (1952/1971).

towards rigorous analysis and consistent proof, but rather as a result of the very logical and set-theoretic contradictions it met with its attempts to press the outer-limits of mathematical comprehending to the comprehensive *per se*. Both the paradoxes and the limitative results were discovered, in fact, on account of the interrogatory momentum and rigor emergent precisely from those critical impulses themselves. Revealingly, Emil Post writes in 1936 of such an critical compulsion as a "psychological fidelity," to "a fundamental discovery in the limitations of the mathematicizing power of Homo Sapiens," a discovery, "in need of continual verification" (1936/1965, p. 291).

The revelation of the antinomies thus in no way dimmed the ethical-practical beam of illumination strengthening from within the deepening normative character this particular sort of shared human practice of rigorous consensing even amidst the stygian shadows of self-reference. In this latter respect, notwithstanding its weakness for the siren's call coming from those suspect figures that populated Cantor's paradise, the challenge posed by Hilbert's program can be viewed as an epistemologically pragmatic challenge to mathematical communication, a challenge having to do with how, given their application of the finite resources of eidetic individuals, groupings and basic operators, claimants could bind each other inferentially through the conveyable coherence, translucence and public-accessibility of the axiomatizations and proofs they offered.

With this in mind, it is perhaps easier to see just how the thinkers of '36 cast a floodlight on the problem of effective computability by bringing it closer to its native normative-practical locus. As the most elemental form of mathematical reasoning, calculation is, in Gödel's absolute sense, reckonability. It is, as the double meaning the English verb "to reckon" indicates, simultaneously a generalizable practice of iterating, of carrying out a combinatorial process of calculation, and a basic way of forming a conclusion for which one is discursively liable, of reckoning in a second, communicatively ramifying sense. It is a way of counting in being held accountable, a way of reckoning in being subject to a continuous tribunal of the reckoning of others. In sum, then, inasmuch as the luminosity of the heroic generation of Turing, the generation of the unsolvable, can hence be seen as deriving from its insights into the moral-practical animation of the social practice of computing and the ethical inner-logic of the incompletability of mathematics, it lives on to today as an Enlightenment that casts no shadows – an Enlightenment not lit from the Platonic heavens above but one rather illumined from between, amidst living human beings who speak, interact and calculate together.

References

Agamben, G. (2009). *What is an aparatus?* (D. Kishik & S. Pedatella, Trans.). Stanford: Stanford University Press.

Apel, K.-O. (1998). *Towards a transformation of philosophy* (G. Adey & D. Frisby, Trans.). Milwaukee: Marquette University Press.

Black, M. (1946). *Critical thinking*. New York: Prentice-Hall.

Brandom, R. (1994). *Making it explicit*. Cambridge: Harvard University Press.
Brandom, R. (2000). *Articulating reasons*. Cambridge/London: Harvard University Press.
Brandom, R. (2009). *Reason in philosophy*. Cambridge: Harvard University Press.
Church, A. (1937). Review of a. m. turing. on computable numbers, with an application to the *Entscheidungsproblem*. *Journal of Symbolic Logic, 2*(1), 42–43.
Dewey, J. (1960). *The quest for certainty*. New York: Penguin Books.
Foucault, M. (1966/1994). *The order of things*. New York: Vintage Books.
Gadamer, H. G. (1960/1992). *Truth and method* (D. Marshall & J. Weinsheimer, Trans.). New York: Crossroad.
Gandy, R. (1988). The confluence of ideas in 1936. In R. Herken (Ed.), *The universal turing machine: A half-century survey* (pp. 55–111). Oxford/New York: Oxford University Press.
Gödel, K. (1931/1965). On formally undecidable propositions of the principia mathematica and related systems. In M. Davis (Ed.), *The undecidable* (pp. 4–38). New York: Raven Press.
Gödel, K. (1946/1965). Remarks before the Princeton bicentennial conference on problems in mathematics. In M. Davis (Ed.), *The undecidable* (pp. 84–87). New York: Raven Press.
Gödel, K. (1951/1995). *Unpublished philosophical essays* (F. A. Rodriguez-Consuegra, Ed.). Basel/Boston: Birkhäuser.
Gödel, K. (1986). *Collected works i* (S. Feferman, J. W. Dawson, S. C. Kleene, G. H. Moore, R. M. Solovay, & J. van Heijenoort, Eds.). Oxford: Clarendon Press.
Goodman, N. (1960). The way the world is. *The Review of Metaphysics, 14*(1), 48–56.
Habermas, J. (1992). *Postmetaphysical thinking* (W. Hohengarten, Trans.). Cambridge: MIT.
Haugeland, J. (1998). *Having thought*. Cambridge: Harvard University Press.
Hofstadter, D. (1979/1994). *Gödel, Escher, Bach: An eternal golden braid*. New York: Basic Books.
Jaspers, K. (1953). *The origin and goal of history* (M. Bullock, Trans.). New Haven: Yale University Press.
Kleene, S. C. (1952/1971). *Introduction to metamathematics*. Groningen: Wolters-Noordhoff.
Kleene, S. C. (1987). Reflections on Church's thesis. *Notre Dame Journal of Formal Logic, 28*(4), 490–498.
Kleene, S. C. (1988). Turing's analysis of computability, and major applications of it. In R. Herken (Ed.), *The universal turing machine: A half-century survey* (pp. 17–54). Oxford/New York: Oxford University Press.
Kripke, S. (2013). The Church turing 'thesis' as a special corollary. In J. Copeland, C. Posy, & O. Shagrir (Eds.), *Computability: Turing, Gödel, Church, and beyond.* Cambridge: MIT.
Lewis, C. I. (1929). *Mind and the world order*. New York: Dover Publications.
Murawski, R. (1999). *Recursive functions and metamathematics*. Dordrecht/Boston: Kluwer Academic.
Post, E. (1936/1965). Finite combinatory processes: Formulation i. In M. Davis (Ed.), *The undecidable.* Hewlett: Raven Press.
Rorty, R. (1982). *Consequences of pragmatism*. Minneapolis: University of Minnesota Press.
Searle, J. (1968). Austin on locutionary and illocutionary acts. *The Philosophical Review, 77*, 405–424.
Sellars, W. (1956/1997). *Empiricism and philosophy of mind*. Cambridge: Harvard University Press.
Sieg, W. (1994). Mechanical procedures and mathematical experience. In A. George (Ed.), *Mathematics and mind.* New York: Oxford University Press.
Sieg, W. (2002). Calculations by man and machine. In W. Sieg, R. Sommer, & C. Talcott (Eds.), *Reflections on the foundations of mathematics* (pp. 396–415). Urbana: Association for Symbolic Logic.
Sieg, W. (2006). On mind and Turing's machines. *Natural Computing, 6*(2), 187–205.
Turing, A. M. (1936/2004). On computable numbers, with an application to the *Entscheidungsproblem*. In B. J. Copeland (Ed.), *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life* (pp. 58–90). Oxford: Clarendon Press.

Turing, A. M. (1938/2004). Systems of logic based on ordinals. In B. J. Copeland (Ed.), *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life* (pp. 146–204). Oxford: Clarendon Press.

Turing, A. M. (1948/2004). Intelligent machinery. In B. J. Copeland (Ed.), *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life* (pp. 395–432). Oxford: Clarendon Press.

Turing, A. M. (1950/2004). Computing machinery and intelligence. In B. J. Copeland (Ed.), *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life* (pp. 433–464). Oxford: Clarendon Press.

Turing, A. M. (1951/2004). Can digital computers think? In B. J. Copeland (Ed.), *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life* (pp. 476–486). Oxford: Clarendon Press.

Turing, A. M. (1954/2004). Solvable and unsolvable problems. In B. J. Copeland (Ed.), *The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life* (pp. 576–596). Oxford: Clarendon Press.

Chapter 6
Is There a Role for Computation in the Enactive Paradigm?

Carlos F. Brito and Victor X. Marques

Abstract The main contribution of this paper is a naturalized account of the phenomenon of computation. The key idea for the development of this account is the identification of the notion of syntactical processing (or information processing) with the dynamical evolution of a constrained physical process, based on the observation that both evolve according to an arbitrary set of rules. This identification, in turn, revealed that, from the physical point of view, computation could be understood in terms of the operation of a component subdivided into two parts, (a) the constrained process and (b) the constraints that control its dynamics, where the interactions with the rest of the system are mediated by configurational changes of the constrained process. The immediate consequence of this analysis is the observation that this notion of computation can be readily integrated into the enactive paradigm of cognition.

Keywords Enaction • Computation • Syntax • Arbitrariness • Computation • Varela • Organism • Searle

6.1 Introduction

The enactive paradigm arises as an alternative to the computationalist program for the cognitive sciences. It emerges at the convergence of the dynamical and the embodied approaches to cognition, proposing to understand the cognitive phenomena from a biological perspective, and pointing to the continuity of life and mind (Varela 1992). From the beginning, the new paradigm shows itself very critical to what Varela calls "the Gestalt of the computer" – a tendency to see the computer as a privileged metaphor in terms of which everything else is measured

C.F. Brito (✉)
Computer Science Department, Universidade Federal do Ceará, Fortaleza, Brasil
e-mail: carlos@lia.ufc.br

V.X. Marques
Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brasil
e-mail: victorxis@gmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_6

(Varela 1979). In opposition to the traditional cognitive sciences, Varela wants to call attention to the disanalogies between cognition and computation, and between brains and digital computers. For him, cognition is co-extensive with life and has to do, not with representing the world, but with maintaining a precarious identity in the face of a constant lack (of materials and energy) and irritations provoked by the environment. In living systems, it is crucial to apprehend how material and energetic flows are constrained in a way to reconstruct the very constraints that control them. Accordingly, the enactive paradigm insists that cognition must be understood as an integral part of this circular dynamics of self-maintenance. Such theoretical compromises naturally lead the enactive paradigm to have a preference for operational descriptions of a causal nature, instead of the symbolic and informational descriptions favored by the computationalist approach.

But, there is also a sense in which the living organisms are the first informational systems in nature. In addition to matter and energy flows, organisms also process (informational) patterns in order to improve their adaptive relation with the environment. Recognizing patterns in the environment and using them to modulate behavior, allow living organisms to be anticipatory systems, which can prepare for and act upon what has not yet happened (Rosen 1985). As observed by Hopfield, "Much of the history of evolution can be read as the evolution of systems to make environmental measurements, make predictions, and generate appropriate actions" (Hopfield 1994, p. 56). Such systems are physical structures with precisely ordered constraints, whose sensitivity and specificity make it possible that trivial physical changes, negligible at the scale of magnitude of the organism (for instance, a change in the rhythm of an interaction), generate significant consequences for behavior. Now, it is clear that a description that concentrates on the physical laws and microscopic details hides the organization of a complex system – which is then uncovered by a structural abstraction. In a similar way, the emphasis on operational descriptions does not allow us to see clearly how the behavior of certain complex systems is organized in terms of the detection and processing of flows of pattern changes. The alternative here is to switch to a different mode of description formulated in terms of informational tokens deprived of causal power and specialized modules that process these tokens. Indeed, Hopfield complements the observation reproduced above with the following remark: "This is an example of computation in the sense that the term is generally understood in computer science".

The main difficulty in accommodating together the two views of the organism presented above (and reconciling the corresponding approaches to cognition) resides in the fact that they are based on two very different modes of description: the so-called operational mode of description, which explains the behavior or functionality of a system by producing a "structural abstraction" and then pointing out to the causal relationships among the components (e.g., consider the typical description of a machine or an explanation of the bodily functions of an animal); and the so-called informational mode of description, which explains the behavior or functionality of a system by producing a "syntactical abstraction" based on representations

and pointing out to modules that perform manipulations on these abstract entities (e.g., consider the typical description of a computer program or an explanation of the cognitive capacities of the brain). So, the problem of the compatibilization of the enactive paradigm and the computationalist program to cognition hinges upon the more fundamental problem of the relation between information and causation.

Our strategy to tackle the problem is based on the formulation of a naturalized account of the phenomenon of computation, aiming at an approximation with the descriptive framework of the enactive paradigm. We begin our exposition with a very brief review of some arguments by Searle and Haugeland on computation, syntax and physics, which served as the initial motivation for this work. Then, we introduce a key concept that will be extensively used in our discussion: the concept of constraint (Pattee 1971; Umerez and Mossio 2013). This concept allows us to formalize the idea of a physical system in which one part controls the dynamical evolution of another part. We capture this idea in the form of a simplified model of an organized physical system: a system of coordinated constraints that controls the evolution of underlying dynamical processes. On the basis of this simplified model, we are then able to formulate precise definitions for the notions of 'structural abstraction' and 'syntactic abstraction' of an organized physical system. As it turns out, the specification of a syntactical abstraction depends on a choice that is not forced by the internal structure of the physical system, and reflects the point of view of an external observer. So, an immediate consequence of our definitions is the elucidation of the main intuition behind Searle's arguments in Searle (1990). The last step towards the naturalization of the phenomenon of computation consists of the derivation of the notion of informational interaction from the notion of syntactical abstraction, by substituting an interacting physical system for the intentional observer. It is important to note that an informational interaction is a regular physical interaction, with specific characteristics that seem to capture some key aspects of the operation of informational systems. At this point, we are finally ready to formulate our naturalized account of computation as *any function performed by a component that relates to the other components of a larger system (or the environment) exclusively through informational interactions*. Notice that this definition focus on the way through which a computer relates to the other components of a physical system, and not on its internal organization or dynamics (e.g., if it is digital or analog). In particular, we are interested in the sort of interactions that define the notions of input and output to a physical system (an issue that is taken for granted in most analyses of computation Piccinini 2007; Scheutz 1999). The final part of the paper is dedicated to the investigation of the problem of the compatibilization of the enactive paradigm with the concept of computation, using the notions that have been introduced. We conclude with some considerations regarding the limitations of the naturalized account of computation that was proposed, and present some indications about how they could be overcome.

6.2 Naturalized Account of Computation

In Searle (1980, 1990) we find several interesting observations about what we mean when we say that a physical system is a digital computer. Taking the fact that "[computer] programs are defined purely formally or syntactically" as his starting point, all the effort by Searle in the chinese room argument is devoted to prove that "syntax is not the same as, nor is it sufficient for, semantics". In the later work, however, Searle moves one step further and claims that "syntax is not intrinsic to physics" and that "syntax is essentially an observer relative notion". In this sense, he argues, a syntactical description cannot be considered an abstraction of a physical system because something which is not there is added by an external observer. Accordingly, he gives the following characterization of digital computers: "To find out if an object is really a digital computer (...) we just have to look for something that (...) could be used to function as O's and 1's" (Searle 1990, p. 6). The point here is that something is a digital computer because someone assigns a syntactical interpretation to the purely physical features of the system. This line of reasoning leads him to such odd conclusions as "for any object there is some description of that object such that under that description the object is a digital computer" (Searle 1990, p. 7). Such conclusions have stimulated many commentators to analyze his arguments, most aiming at refuting them.

John Haugeland criticizes Searle's point of view arguing that his analysis is not precise enough in the sense that it leaves some essential features of the concept of computation out (Haugeland 2002). He points out that, besides being describable in purely syntactical terms, it is also essential to computer programs that "there must be possible concrete implementations of them". That is, a physical system whose causal interactions reliably correspond to the operations on the data that are prescribed by the program. So, to the extent that these operations correspond to modifications on concretely implemented data structures, "carrying out operations must be a causal process". From these considerations, Haugeland concludes that the purely syntactical description can indeed be regarded as an abstraction from the (possibly multiple) concrete implementations of the program. As he puts it: "this is exactly like the relation between engineering drawings (...) for a pump (...) and the various possible actualizations of such pumps" (Haugeland 2002).

Actually, these two types of abstraction (the computer program and the engineering drawing) are not identical, and Haugeland introduces yet a third essential feature of computer programs that allows him to distinguish between them. The idea is that programs must also be describable *semantically*. Programs are written in a general code which permits one to specify (or prescribe) arbitrary operations to be performed on the data, and the processor (hardware) must be capable of carrying out whatever operation is prescribed. The general code typically has a compositional structure, with a semantics associated with individual program tokens (e.g., terms for arithmetic operations), which are then combined to induce the semantics of entire expressions. In this sense, the program tokens can be regarded as symbols – that is, they are *meaningful* – and Haugeland works to

make the point that they are meaningful to the processor itself: "the processor responds to [the program instructions] as prescriptions with those semantics", and "The only way that we can make sense of a computer as executing a program is by understanding its processor as responding to the program prescriptions as meaningful" (Haugeland 2002). So, the syntactical description of a computational system (i.e., the program) is incomplete by itself not only because it presupposes a concrete causal implementation, but also because the semantics of the program is lacking. Haugeland finally concludes that syntactical descriptions of computer programs are fundamentally different from engineering drawings of pumps because the later "do not specify any part of the pump (...) as itself explicitly prescribing what a second part of it is to do to a third part" (Haugeland 2002).

6.2.1 Constraints and Organized Physical Systems

According to the Encyclopedia of Systems Biology, "constraint refers to a reduction of the degrees of freedom of the elements of a system exerted by some collection of elements, or a limitation or bias on the variability or possibilities of change in the kind of such elements" (Umerez and Mossio 2013). The simplest illustration of the concept is given by a mechanical physical system where the constraint plays the role of a boundary condition. For example, the trajectories of the molecules of a gas inside a container are constrained by the walls of the container. In a different sort of example, strings joining pairs of balls moving in a billiard table (or the chemical bonds in a molecule) define rigidity constraints that reduce the number of dynamical variables in the system and create structure. Finally, catalyzers accelerating reactions in a chemical solution provide an example of a constraint that introduces a bias on the possibility of change in the kinds of elements in the system.

This general definition captures a common intuition about constraints as something that defines boundaries, or extreme points for variation, but does not in itself determine specific behavior. According to this understanding, behavior is still governed by the laws and properties of the elementary parts of the system (physical, chemical or otherwise). But, as indicated further on in the article of the Encyclopedia (Umerez and Mossio 2013, p. 491), more specific characterizations of the concept of constraint have been formulated in several domains (Ashby 1958; Pattee 1971; Polanyi 1968) in attempts to ground explanations using concepts of the physical sciences.

A brief search in the literature shows that the term constraint is used, somewhat ambiguously, with several related meanings: sometimes it refers to a material structure that embodies the constraint, sometimes it refers to the effect of such structure on the dynamics of the other elements of the system, and sometimes it is used as a technical term for a tool available to the physicists to produce simpler explanations or predictions of the behavior of the system. In all cases, however, we are in the context of an alternative description of a physical system in which we describe the dynamics of one part of the system constrained by another part which

is typically held fixed.[1] In technical terms, this alternative description is obtained by introducing additional equations of constraint into the usual set of dynamical equations of the system. As a simple example, when the physicists describe the experiment of a ball rolling down an inclined plane, they do not consider the detailed forces of interaction between the ball and the plane, but just introduce a new equation in their model to enforce that the trajectory of the ball will obey the constraint (i.e., the inclined plane). This equation, of course, does not take into account the microscopic details of the particles that compose the plane, but simply describes a geometric plane at a position that corresponds approximately to the average position of the surface particles of the actual plane. This selective omission of degrees of freedom of the system naturally leads to a great simplification in predictions and explanations.

Still following the Encyclopedia, the notion of constraint is usually employed in relation to conceptualizations in terms of levels or hierarchies. More specifically, the concept of constraint is used to give precise expression to the idea of interactions between different levels of organization. This is an important point for our discussion that deserves further clarification. In the simplest case of artificially constructed systems, the constraints are typically produced and maintained by processes at higher levels of organization (e.g., human activity and design), and they affect the dynamics of the lower level elements of the system in very specific ways. Living organisms, on the other hand, are autonomous systems in the sense that they produce their own constraints; that is, the constrained dynamics of the lower level elements of the system results in the maintenance and replacement of the very constraints that control this dynamics, in a closed loop. In both cases, the specific *structure* or *form* of a given constraint is not the result of the dynamics of the immediate elements that it constrains.[2] In this sense, from the perspective of the constrained lower level processes, the effect of the constraint on the dynamics corresponds to an arbitrary external intervention. Finally, if we recall that the constraint is associated with the dynamics at a higher level of organization, then we get the idea of control.

Now, what is already clear in this picture is that the role of the constraints here is not anymore simply that of establishing arbitrary limitations to variation, but they actually participate in the determination of specific behavior. That is, in the context of control, the dynamics of the lower level elements of the system is effectively governed *both* by their specific laws and properties and by the actuation of the constraints. It is precisely in this sense that M. Polanyi characterized machines and organisms as dual control systems, and used the suggestive image of constraints (or, boundary conditions, in his terms) "harnessing the laws of nature" in order to produce work (Polanyi 1968).

[1] See Rosen (1986) for a more precise characterization of the modes of description offered by Newtonian particle mechanics and analytic mechanics, where the concept of constraint was first introduced.

[2] It should also be mentioned that, in the more interesting cases, the *operation* of the constraint, i.e., the particular way in which the constraint affects the lower level elements, may depend on the state of those elements at a given moment.

The natural next step, is to consider physical systems whose detailed state or configurational dynamics is completely controlled by constraints, and energy considerations do not enter in the determination of specific behavior. Robert Rosen (1986) designates such systems as maximally (or totally) constrained systems and provides a formal characterization for them in terms of non-holonomic constraints (i.e., a constraint that removes a velocity degree of freedom, but leaves the dimensionality of the configurational space unchanged). Due to space limitations, we cannot reproduce Rosen's analysis here, but for the purposes of our discussion the following quotation should suffice: "If we impose the maximal number of non-holonomic constraints, then the velocity vector is uniquely determined by the configuration. The result is an autonomous dynamical system, or vector field, on the configuration space. At this point the impressed forces of conventional analytic mechanics disappear completely; their only role is to get the system moving. Once moving, the motion is completely described by the constraints" (Rosen 1986, p. 112).

Rosen defines the notion of maximally constrained systems in the context of mathematical models, where the physical systems is defined by the collection of dynamical variables that appear in the equations. Here, we want to keep the idea that there is one part of the system (the constraints) that controls the dynamical evolution of another part (the constrained processes). For this reason, we will make use of the weaker notion of a system which is maximally constrained only with respect to a subset of its dynamical variables. That is, the constraints of the system strictly control the configurational dynamics of a number of variables associated with the constrained processes of the system.

We summarize the discussion of this section with the definition of a simplified model of an organized physical system:

Definition 6.1. An *organized physical system* is a system defined by a number of higher-level constraints whose coordinated action controls the dynamical behavior of underlying physical processes inside the system.

6.2.2 Structural and Syntactical Abstractions

It is already well recognized that organized physical systems, such as organisms, machines, and even some dissipative structures like a candle flame, cannot be properly described in terms of collections of particles following trajectories governed by the fundamental laws of physics (Mossio et al. 2009; Polanyi 1968; Rosen 1991; Varela 1979). Such descriptions completely miss what is most relevant about the system: its organization. A better description of such systems would present only the main structures that define their higher-level constraints, as well as the relations that establish the coordination among those constraints. In this type of structural or relational description, all the emphasis is placed on the high-level physical organization of the system.

Definition 6.2. A *structural abstraction* of an organized physical system is a description which presents the structures that define the higher level constraints of the system, as well as the relations that establish their coordination, but omits most (or all) the information about the dynamical processes constrained by those structures.

As we said in the beginning, this is already well understood. What we propose here is that, under appropriate circumstances, an organized physical system can also be the object of a syntactical description. The paradigm we shall adopt for syntactical descriptions is the notion of a computer program. Our basic idea is to associate the execution of the program with the dynamics of the constrained processes of the organized physical system. We begin with the observation that the execution of a computer program corresponds to a sort of flow in which informational elements are transformed by abstract operations and get involved in abstract relationships (e.g. comparison). So, the first difficulty that confronts us is the fact that the execution of the program does not correspond to a flow of energy or matter. What is being described is a flow of (pattern) changes in time. To make the correspondence work, then, we will associate the flow of the program (i.e., the computation) with the changes of configuration suffered by specific dynamical variables associated with the constrained processes of the organized physical system.

In order to make things concrete, consider the example of the electronic computer. We usually describe its computation from the point of view of information processing, using the abstract and more convenient language of numbers and strings of symbols. However, a computation in the physical device consists of a sequence of changes of voltage levels at specific locations of the machine, controlled by the pieces of electronics (constraints) that constitute the hardware. But, perhaps, our familiarity with the example does not allow us to see the point fully clear. What we want to say is that, in principle, the "syntactic" point of view is always available to describe what is going on in an organized physical system. All that is required is a choice of dynamical variables with respect to which the system is maximally constrained. Once this choice is fixed, we simply describe how the variables change in time under the action of the constraints. Since the configurational dynamics is strictly controlled by the constraints, the description can be provided in the form of rules that update the configuration of the selected variables.

Definition 6.3. A *syntactical abstraction* of an organized physical system is a description that presents the rules that govern the changes of configuration of a number of selected dynamical variables of the system, as well as the relations that coordinate the application of these rules, but omits all the information about the structures (or constraints) that implement the rules and the coordination.

This mode of description is remarkably general. The dynamical variables which are used to define the syntactical abstraction can be just about anything (water levels, sound frequencies, spatial configurations of mice and cheese, etc.), as witnessed by the variety of strange 'computers' described in the literature, which were contrived to illustrate the independence of the phenomenon of computation with respect to the material substrate (Block 1995).

6.2.3 *Naturalizing the Observer*

The discussion in the previous section exposed the objective aspects which are inherent to any syntactical description of an organized physical system: the structural constraints that implement the rules that appear in the description. Nevertheless, the analysis also made it clear that there exists a subjective residue that cannot be eliminated, in the sense that it cannot be accounted for exclusively in terms of aspects of the system itself: the choice of the dynamical variables whose behavior is the target of the description. This means that it only makes sense to talk about syntactical descriptions in the actual presence of an observer that makes this choice. In other words, to obtain a naturalized account of the phenomena of computation we must explicitly include the observer in the analysis. In order to do so, we have to abandon the idea of an observer as an intentional entity which contemplates the system and eventually produces a description. Instead, we will assume that the observer is another physical system which engages into a causal interaction with our original organized system.

The first step that must be taken in order to make the move from description to interaction is to check whether the language of organized systems can still be used to describe the interaction between two physical systems. That is, we have to see if the notions of higher-level constraints, lower-level details, underlying constrained processes, etc., are not just convenient constructs invoked by an intentional observer in order to simplify her description. But, the whole point of Sect. 6.2.1 was that the notion of constraint corresponds to the objective fact that some degrees of freedom are, for all practical purposes, eliminated from the system due to the presence of strong forces and/or material structures which are themselves not affected by the dynamics. This point is at the root of the concept of constraint formulated by Howard Pattee (1971).

Next, we consider the following question: what sort of interaction corresponds to a syntactical description? If we denote our original organized physical system by C (the computer) and denote the system we have just introduced in the story by O (the observer), then we should expect this to be an interaction between the observer system O and the constrained processes of the computer system C. But we need to proceed carefully here. Recall that the role of the observer is just to choose the dynamical variables whose behavior is the object of the syntactical description. The rules that govern this behavior should be implemented by the structural constraints of the computer system. We impose two requirements in order to satisfy these conditions. First, the interaction between the observer system O and the computer system C must be mediated by dynamical variables associated with the constrained processes of the computer component C. That is, the behavior of O should be sensitive to state changes of these variables but, once this state is fixed, variations of other aspects associated with C should have no effect on the behavior of O. Second, the computer system C must be maximally constrained with respect to the variables thus selected by O.

These requirements have the following important implication: the energy which is eventually exchanged between the systems C and O is irrelevant from the point

of view of their interaction. To see this, we just need to recall the definition of a maximally constrained system, which implies that the configurational dynamics of the mediating variables is completely determined by the constraints of the computer system. On the other hand, the observer system is only affected by the configuration that is produced by the computer system. An evidence that this should indeed be the case, is confirmed by our familiar experience with electronic computers, where it doesn't matter how hard we press the buttons when we type the input. This motivates us to characterize the interaction between C and O as an informational interaction, in the sense that the interaction is not to be explained in generally mechanical terms (i.e., how the two systems push and pull each other).

Definition 6.4. An *informational interaction* is a causal interaction between a physical system O and an organized physical system C such that: (a) the interaction is mediated by the changes of configuration of a number of dynamical variables associated with the constrained processes of C; (b) the organized system C is maximally constrained with respect to the variables that mediate the interaction.

6.2.4 Naturalized Computation

In Haugeland (1981), J. Haugeland characterizes a computer as an automatic formal system. That is, a system defined in terms of a number of rules that change the configuration of an underlying set of variables, endowed with an intrinsic dynamics that controls the application of those rules without external interference. It is clear that this notion can be captured by the developments of the previous sections: (1) the idea that an organized physical system can be the object of a syntactic description allows us to view it as a formal system; (2) an interaction mediated by the changes of configuration of some variables associated with the organized system, defines a notion of input-output that gives access to the operation of the formal system; (3) if the organized system is maximally constrained with respect to the variables that mediate the interaction, then it also possesses the autonomous behavior required by the notion of automatic formal system. These three conditions are encapsulated in the definition of informational interaction.

However, the notion of informational interaction places almost no restriction on the observer system. It just says that there is a causal interaction between the observer system and the computer system, mediated by the changes of configuration of some variables associated with C. But, there are many irrelevant ways in which one physical system can causally affect another one. So, if we do not qualify this interaction, then we take the risk of postulating that there is computation in many cases where there is actually not. The obvious idea here is that the interaction with the computer system C should be relevant for the activity of the observer system. Indeed, in Haugeland (1981), Haugeland also says that, in addition to being an automatic formal system, a computer is also a semantic engine. That is, a system

which somehow relates with the external world, and whose overall behavior and specific workings can be interpreted or attributed meanings in terms of the objects and events of this world.

In order to capture this additional aspect of computers, and to avoid the accidental cases of computation, we introduce the last element of our naturalized account of computation: (4) the computer system C must have a function in the activity or operation of the observer system O. Now, with this choice of using the concept of function we are committing ourselves to talk about computation only in the contexts in which this sort of teleological discourse applies – that is, the context of living organisms and their tools – which is quite reasonable. The concept of function is largely unproblematic in the social and technological context in which we construct and use our computers. On the other hand, to obtain a more general notion of computation associated with natural phenomena, we can make use of the naturalized accounts of functions presented in Mossio et al. (2009) and Bickhard (2000), according to which a component has a function in the context of the operation of an autonomous system (i.e., an organism) if the component contributes to the self-maintaining activity of the system.

Definition 6.5. *Computation* is any function performed by a component which relates to the other components of a larger system (or the environment) exclusively through informational interactions.

6.3 Computation in the Enactive Paradigm

Now, we return to the problem of the compatibilization of the enactive paradigm with the notion of computation. As we mentioned in the Introduction, the main difficulty lies in the fact that we tend to understand computation abstractly, in syntactic terms, while the enactive paradigm is concerned with the embodied experience of an agent in the world. Actually, this is basic problem of the cognitive sciences: to deal with the apparent incompatibility between the abstract and the concrete.

The computationalist approach to cognition proposes to solve the problem by postulating that the brain processes information received from the sensory organs and outputs instructions to the body which executes them. However, it has never been clear how the sensory embodied experience becomes information, and how, after processing, the information becomes bodily behavior. All the effort was concentrated on understanding the mechanisms behind the information processing. The problem presented by the 'translations' was relegated to be solved by interfaces, which were considered unimportant and not at the center of the phenomenon of cognition. This solution worked well for a while in the AI community, because in practice the human beings were playing the role of the interfaces: coding the (relevant) facts of the physical reality into informational tokens, and interpreting the results of the computation by taking the appropriate actions in the world. Also, in

very simplified settings, the computer could be put in direct contact with reality, but here again the problem was solved by a human being, who designed the appropriate interfaces. Soon it became clear that there was an important difficulty here.

The solution offered by the enactive approach consists of leaving computation out and interpreting the immediate interactions of the living cell with the elements of its environment, through adaptive mechanisms, as cognition (Varela 1992). The paradigmatic example is given by the bacteria swimming up a sucrose gradient. To understand the example it is important to note the distinction between the environment as it appears to an observer and without reference to an autonomous unity, and the environment *for* the living system. The molecule of sucrose only acquires significance (meaning) as food in the presence of the autopoietic activity of the bacteria, which makes use of this molecule to continue its process of self-maintenance. Varela calls this a surplus of significance, created by "the presence and perspective of the bacteria as a totality", and he claims that this is the mother of intentionality. It is important to call attention to the two components of this definition. The first is the fact that the specific meaning acquired by the molecule of sucrose (i.e., food) is determined by the particular way in which the molecule is integrated in the autopoietic (and adaptive) activity of the bacteria. The second component is the fact that this activity is a manifestation of an autonomous entity, so the attribution of meaning is not something that depends on the judgement of an external observer. Here, it is useful to quote Varela again: "what is meaningful for an organism is precisely given by its constitution as a distributed process, with an indissociable link between local processes where an interaction occur, and the coordinated entity which is the autopoietic unity" and "this permanent relentless action (...) becomes, from the observer side, the ongoing cognitive activity of the system" (Varela 1992, p. 8).

However, Moreno et al. (1992) highlight an important difference between purely adaptive systems and properly cognitive ones. They say that while "ontogenetic adaptation ensures, through perception, the functional correlation between metabolic-motor states and states of the environment" (Moreno et al. 1992, p. 66), cognition is related to learning, memory and anticipatory behavior, and those involve "the capacity (...) to change, in somatic time, the very structure of the system that correlates sensors and motors" (Moreno et al. 1992, p. 68). This observation exposes a serious limitation of the enactive approach to explain higher level cognition: according to the enactivism, the meaning of an element is associated with the way through which it is integrated in the autopoietic and adaptive activity of the living system, but this is not something easy to change (or create) in somatic time – think of the intricate network of relations which underlies the metabolic activity of the living cell. Here, the notions of information and computation become very attractive – by the usual definition, computation is a process of manipulation and transformation of information. So, if the living system acquires the capacity of performing computation, then the problem of cognition is solved (from the biological point of view). Indeed, Moreno et al. associate cognition with "a specialized subsystem continuously reconstructing patterns that are functional or referentially correlated

with changes in the environment" (Moreno et al. 1992, p. 67). But this brings back all the problems associated with the computationalist approach to cognition, and again they are relegated to be solved by appropriate (but unspecified) interfaces, as is clear from the following passage "in cognitive organisms, the physical patterns impinging on sensors are translated in trains of discrete sequences that modify the dynamics of a network of information processing" (Moreno et al. 1992, p. 67).

We learn two things from this discussion: (1) that it is possible to formulate a consistent naturalized account of meaning and cognition in the context of the activity of an autonomous system, as long as only mechanical and chemical interactions are involved; and (2) that we cannot dispense with some notion of information and/or computation if we want to keep the goal of explaining higher level cognition. The problem is how to accommodate the notion of information/computation and the notion of purely natural systems in the same explanation. This difficulty is well synthesized by Searle when he exclaims that "syntax has no causal powers". Here is the place where we can apply the insights of our naturalized account of computation.

The first thing to note is that there is no difficulty in extending the enactive account of meaning and cognition to the situation of an autonomous system with a component that performs computation, as defined in Sect. 6.2.4. To see this, consider an autonomous system O with a computer component C that: (a) has a function in the self-maintaining activity of O; (b) relates to the other components exclusively through informational interactions. Here, we recall that an informational interaction is just a specific form of causal interaction between physical systems. This means that we can understand the operation of the autonomous system O in terms of the causal relationships that hold among its components. Moreover, by focusing on the structural constraints of the computer component, we see that they are coordinated with the constraints of the other components and, in this way, they contribute to control the behavior of the autonomous system. The only specificity here is that this coordination is established through the precise manipulation of the configuration of a number of dynamical variables. From this perspective, we can see that there is nothing going on in this system that does not happen in the living cell – no "vaporous" notion of information (Varela 1997, p. 79) affecting the physical workings of the system. So, the enactive account of meaning also holds in the context of the mechanical, chemical and information interactions (as defined in Sect. 6.2.3) of an autonomous system.

On the other hand, we can also focus our attention on the constrained processes of the component C, and more specifically on the set of variables X that mediate the informational interactions between C and the other components of O. Then, we observe that the component C accomplishes its function by setting the variables X to specific configurations. Hence, it makes sense to interpret the variables X and their particular configurations in terms of the function performed by C. For example, if C has a regulatory function in the self-maintaining activity of O, then the configuration of the variables X before the interaction would be interpreted as an indication of some specific event or state of the autonomous system O, and the

configurations produced by the component C would be interpreted as instructions for other components to take appropriate actions. Next, we recall that C is maximally constrained with respect to the variables *X*. This means that the dynamics of changes of configuration of the variables *X* is strictly controlled by the constraints of the component C and, in this sense, it can be described as governed by a set of arbitrary rules. In other words, we can explain how the component C performs its function in syntactical terms. So, from this perspective, it is reasonable to say that the component C is doing computation, in the usual sense of the term.

Returning to the issue of higher-level cognition, Moreno and Lasa (2003) again associate it with a specialized subsystem which is now qualified as *informational* and *dynamically decoupled* from general metabolic processes. The subsystem which they have in mind is the nervous system and what makes it special are the two forementioned properties. The phenomenon of decoupling is explained as follows: a part of the system constitutes a new level of interactions which operates according to a set of arbitrary rules independent of the dynamics of the lower level (the remaining system); (b) both levels become causally connected and depend on each other. On the other hand, they give the following reasons to qualify the activity of the nervous system as informational: (a) neural states can switch body states by configurational rather than energetic means; (b) being dynamically decoupled from metabolic processes, these configurations can recursively operate on themselves, producing a kind of "formal processing".

This description is very close to our naturalized account of computation, but we want to call attention to an important difference. When Moreno and Lasa talk about a part of the system which operates according to a set of independent rules, where configurations recursively operate on themselves, they seem to be referring to (in our terms) the constrained processes of the computer component. However, they do not mention the constraints that control these processes – actually, they do not make an explicit distinction between the two parts of the computer component. Now, what becomes clear from our analysis is that the set of independent rules that governs the computation does not correspond to the intrinsic laws and properties of the elements of the constrained process and their particular configurations. These rules reflect the specific and arbitrary action of the structural constraints of the computer component. In other words, the configurations do not operate on themselves, but they are operated upon according to rules that are physically implemented by constraints. So, an adequate naturalization of the concept of computation should present the computer component not only as dynamically decoupled from its surroundings and operating on configurations, but also as subdivided in two levels: (a) a lower level whose dynamics is described by arbitrary rules that modify an underlying configuration, and which is the part of the subsystem that is (directly) causally connected to the rest of the system; (b) an upper-level which consists of the constraints that physically control the dynamics of the lower level. Now, the puzzle of the causal power of syntactic entities arises from the fact that the descriptions of informational systems typically present the part (a) above but omit the part (b). When the part (b) is also presented, the puzzle disappears.

6.4 Conclusion

The main contribution of this paper is a naturalized account of the phenomenon of computation. The key idea for the development of this account was the identification of the notion of syntactical processing (or information processing) with the dynamical evolution of a constrained physical process, based on the observation that both evolve according to an arbitrary set of rules. This identification, in turn, revealed that, from the physical point of view, computation could be understood in terms of the operation of a component subdivided into two parts, (a) the constrained process and (b) the constraints that control its dynamics, where the interactions with the rest of the system are mediated by configurational changes of the constrained process. Now, once we have such an operational characterization of computation, it is an easy step to show that it can be incorporated into the enactive account of cognition.

On the other hand, it is clear that this account captures only one aspect of our usual concept of computation: the peculiar way in which computers relate to the other components of a larger system. Perhaps it may be argued that this is the relevant aspect to capture if we are interested in natural mechanisms of computation, which may not share the structural organization of the digital computer. In any case, there are two limitations of the present account that we believe could be overcome using ideas related to the ones presented here. The first one has to do with the fact that we did not say anything about how a component can perform a function in a physical system by a mere rearrangement of the configuration of some dynamical variables. The second one is related to the common intuition that computation is essentially connected with the manipulation of discrete symbols, a point that we also did not touch. Both of these issues could be addressed through a naturalized account of the concept of information, and again the main tool to develop it would be the concept of constraint.

References

Ashby, R. (1958). *An introduction to cybernetics*. London: Chapman & Hall.

Bickhard, M. (2000). Autonomy, function and representation. *Communication and Cognition – Artificial Intelligence, 17*, 111–131.

Block, N. (1995). The mind as the software of the brain. In L. R. Gleitman (Ed.), *An invitation to cognitive science*. Cambridge: MIT.

Haugeland, J. (1981). Semantic engines: An introduction to mind design. In J. Haugeland (Ed.), *Mind design: Philosophy, psychology, artificial intelligence* (pp. 1–34). Cambridge: MIT.

Haugeland, J. (2002). Syntax, semantics, physics. In J. Preston & M. Bishop (Eds.), *Views into the Chinese room: New essays on Searle and artificial intelligence* (pp. 379–392). Oxford: Oxford University Press.

Hopfield, J. (1994). Physics, computation, and why biology looks so different. *Journal of Theoretical Biology, 171*, 53–60.

Moreno, A., & Lasa, A. (2003). From basic adaptivity to early mind. *Evolution and Cognition, 9*, 12–30.

Moreno, A., Merelo, J., & Etxeberria, A. (1992). Perception, adaptation and learning. In *Proceedings of a Workshop on Autopoiesis and Perception*, Dublin. http://www.eeng.dcu.ie/~alife/bmcm9401/contents.txt.

Mossio, M., Saborido, C., & Moreno, A. (2009). An organizational account of biological functions. *British Journal for the Philosophy of Science, 60*(4), 813–841.

Pattee, H. (1971). Physical theories of biological co-ordination. *Quaterly Reviews of Biophysics, 4*, 255–276.

Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science, 74*(4), 501–526.

Polanyi, M. (1968). Life's irreducible structure. *Science, 160*, 1308–1312.

Rosen, R. (1985). *Anticipatory systems*. Oxford: Pergamon Press.

Rosen, R. (1986). Causal structures in brains and machines. *International Journal of General Systems, 12*, 107–126.

Rosen, R. (1991). *Life itself: A comprehensive inquiry into the nature, origin and fabrication of life*. New York: Columbia University Press.

Scheutz, M. (1999). When physical systems realize functions. *Minds and Machines, 9*(2), 161–196.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417–457.

Searle, J. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association, 64*, 21–37.

Umerez, J., & Mossio, M. (2013). Constraint. In W. Dubitzky (Ed.), *Encyclopedia of systems biology* (pp. 490–493). New York: Springer.

Varela, F. (1979). *Principles of biological autonomy*. North Holland: Elsevier.

Varela, F. (1992). Autopoiesis and a biology of intentionality. In B. McMullin & N. Murphy (Eds.), *Proceedings of a Workshop on Autopoiesis and Perception*, Dublin City University.

Varela, F. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition, 34*, 72–87.

Chapter 7
Natural Recursion Doesn't Work That Way: Automata in Planning and Syntax

Cem Bozşahin

Abstract Natural recursion in syntax is recursion by linguistic value, which is not syntactic in nature but semantic. Syntax-specific recursion is not recursion by name as the term is understood in theoretical computer science. Recursion by name is probably not natural because of its infinite typeability. Natural recursion, or recursion by value, is not species-specific. Human recursion is not syntax-specific. The values on which it operates are most likely domain-specific, including those for syntax. Syntax seems to require no more (and no less) than the resource management mechanisms of an embedded push-down automaton (EPDA). We can conceive EPDA as a common automata-theoretic substrate for syntax, collaborative planning, i-intentions, and we-intentions. They manifest the same kind of dependencies. Therefore, syntactic uniqueness arguments for human behavior can be better explained if we conceive automata-constrained recursion as the most unique human capacity for cognitive processes.

Keywords Recursion • Syntax • Planning • Mind and computation

7.1 Introduction

One aspect of theoretical computer science that is useful in AI and cognitive science is in making ideas about computing explicit, independent of whether we are computationalist, cognitivist, connectionist, dynamicist, or an agnostic modeler.

One such concept in need of disambiguated use in cognitive science and linguistics is recursion. A one-time conference was dedicated solely to the discussion of the role of recursion in language and cognition (Speas and Roeper 2009). Current work touches on several issues addressed there, such as its role in planning and syntax, and on lack of recursion in the lexicon (which is only true for a certain kind

C. Bozşahin (✉)
Cognitive Science Department, The Informatics Institute, Middle East Technical University, Ankara, Turkey
e-mail: bozsahin@metu.edu.tr

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_7

of recursion). The critical issue, I believe, is lack of agreement in what we think we are observing in the processes that are called recursive.

The need for agreement arises because very strong empirical claims have been made about recursion's role and its mechanism, such as that of Hauser et al. (2002) and Fitch et al. (2005), where *syntactic* recursion is considered the most unique human capacity, the so-called core computational mechanism in "narrow syntax." The claim follows Chomsky's recent theorizing, in particular the Minimalist Program (Chomsky 1995, 2005), which puts a recursive/cyclic merger at its core, not only as a theoretical device but also as an operation of the mind.

We can conceive this syntax-based (and language-centered) argument about cognition in at least two ways. In its first sense we can take the phrase *syntactic recursion* to mean recursion in syntax, which seems to be everybody's assumption,[1] therefore not expected to be problematic. In the second sense we can take it to mean reentrant piece of knowledge, known as recursion by name (or label) in computer science.[2] As will be evident shortly, these two aspects are not the same when we take formal semantics and formal definitions of recursion into account.

The current paper aims to show that either conception of syntactic recursion poses problems for the narrow claim of narrow syntax, and to the so-called generative enterprise (Huybregts and van Riemsdijk 1982), which claim that syntactic recursion is the unique human capacity. The most uniquely human capacity may be recursion of a certain kind, but it is not limited to recursion in syntax, and it is certainly not syntactic recursion in the second sense above, therefore the conjecture of Chomsky and his colleagues is probably too strong and premature.

The following arguments are made in the paper. The last point is raised as a question. Some of these arguments are quite well-known. I will be explicit about them in the text.

(1) a. Natural recursion in syntax, or recursion by linguistic value, is not syntactic in nature but semantic.

[1] All but one, that is. Everett (2005) argues that recursion is not a fact for all languages. That may be true, but the fact remains that some languages do have it, and all languages are equally likely to be acquirable. See Nevins et al. (2009) and Bozsahin (2012) for some criticism of Everett, and his response to some of the criticisms (Everett 2009). Even when syntactic recursion is not attested, there seems little doubt that semantic recursion, or recursion by value, is common for all humans, e.g. the ability to think where thinker is agent and thinkee is another thought of same type, manifested in English with complement clauses such as *I think she thinks you like me*. But, it can be expressed nonrecursively as well: *You like me. That's what she thinks. That's what I think*. We shall have a closer look at such syntactic, semantic and anaphoric differences in recursive thoughts.

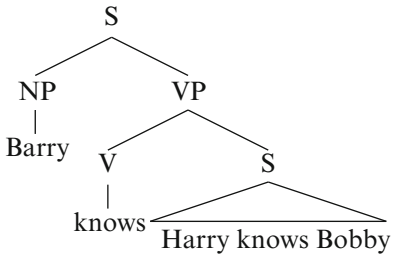
[2] The name is apt because, as lambda-calculus has shown us, reentrant knowledge *can* be captured without names if we want to, and that the solution comes with a price (more on that later). In current work, the term *recursion by name* (or label) is taken in its technical sense in computer science. Confusion will arise when we see the same term in linguistics, for example most recently in Chomsky (2013), where use of the same label in a recursive merger refers to the term 'label' in a different sense, to occurrence of a value.

b. Syntax-specific recursion is not recursion by name as the term is understood in AI and theoretical computer science.
c. Recursion by name is probably not natural.
d. Natural recursion, or recursion by value, is not species-specific.
e. Human recursion is not syntax-specific, although the values it operates on are most likely domain-specific, including those for syntax.
f. Syntax seems to require no more (and no less) than the resource management mechanisms of an embedded push-down automaton (EPDA).
g. We can conceive EPDA as a common automata-theoretic substrate for syntax, collaborative planning, i-intentions, and we-intentions (Searle 1990).
h. The most unique human capacity appears to be the use of recursion with a stack of stacks. Arguments from evolution are needed to see whether planning, syntax or something else might emerge as its first manifestation.

7.2 Recursion by Value Is Semantics with a Syntactic Label

The kind of recursion manifested in language is exemplified in (2), where a sentence (S) contains another sentence as a complement.

(2)



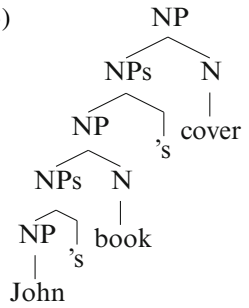
It can be applied unboundedly: *I think you think Barry knows Harry knows Bobby*, etc. The natural capture of this behavior in linguistics is by drawing trees, as above.

It is the property of a class of verbs such as *think, know, ask, claim* that they take such complements. This behavior is constrained by language-particular syntactic properties interacting in complex ways with the argument structure (i.e. semantics) of the event/action/state denoted by the predicate.[3] For example, the English verb *hit* cannot take such complements: **John hits (that) Barry drags Bobby*.

Recursion is possible in the nominal domain as well. For example, a fragment of *John's book's cover's colour* is shown below.

[3]Some comprehensive attempts in linguistics in accounting for the interaction are Grimshaw (1990), Manning (1996), and Hale and Keyser (2002).

(3)



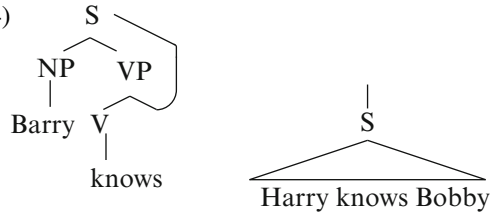
All the grammatical examples above are recursion by value, where *another* instance (i.e. value) of a predicate is taken as an argument of the predicate. For example, *know* takes a knower and a knowee, and the knowee is another predicate: there are two acts of knowing in (2), not one, as the use of same label might suggest, and by two different subjects. The same observations apply to two different possessors and their possession in (3).

The constraints on syntactic values—a better term might be *syntacticized values*—such as S and NPs are semantic in nature, as the distinction between *know* and *hit* shows. The same can be said about the possessive construction: it needs the syntactic correlate of a participant and a property or another participant.

7.3 Syntax-Specific Recursion Is Not Recursion by Name

It follows that the structure below cannot be the right one for *Barry knows Harry knows Bobby*; compare it with the one in (2). In this interpretation, *Harry knows Bobby* would be the base case of recursion by name, as shown separately.

(4)



There is one knower in this structure, and one knowee, but, unlike (2), it points back to the root predicate-argument structure of the word *know*. Both must be *repeated* at every iteration until the base case (the nonrecursive case written on the right) applies to stop the recursion, giving rise to examples such as *Barry knows Barry knows Barry knows Harry knows Bobby*.

It would be tempting to think of this structure as a generalization of (2), by which we would assume (2) to be the base case of (4) without recursion: *Barry*

knows Harry knows Bobby. However, this proposal could not adequately capture the speaker's intuition, that knower can know a knowee, and that, from this she would not infer that knowee's knower must be the same as the knower next level up if multiple embeddings are involved. Neither would she conclude that knower's knowee must be the same predicate until recursion stops, for example *Barry knows John claims Harry knows Bobby*, which is not captured by (4).

It is precisely for this reason that Lexicalized Tree-Adjoining Grammar (LTAG; Joshi and Schabes 1992) represents the reentrancy implied by (2) not as (4) but as an S tree dominating another S tree, one with a special operation of adjunction rather than substitution. Because it is *another* tree, LTAG captures the right semantics of (2). Generative grammar assumes two trees as well, but makes no such combinatory distinction. Therefore, it is susceptible to recursion by name vs. value arguments. Lobina and García-Albea (2009) suggest 'merge' is closure and only 'move' is recursive. The narrower claim here is that any natural recursion must be by value.

The structure in (4) is precisely what is called recursion by name in computer science, considered to be a special form of reentrancy. As the preceding argument shows, it is not the same as recursion by value.

We can have a look at formal definitions of recursion, and also at some recursive definitions, to see what is at stake in deciding what kind of recursion is involved.

Below are two different definitions of a potentially recursive data structure, the tree, from Knuth (1968: 314).

(5) a. Tree: (i) a node called *root* is a tree, denoted as T(root). (ii) The subtrees of a tree T, T($T_1, T_2, \cdots T_m$), are partitioned into $T_1, T_2, \cdots T_m$, where each T_i is a tree.

b. Tree: Any tree is a collection of *nested sets*. A collection of non-empty sets is nested if, given any pair X, Y of the sets, either $X \subseteq Y$ or $X \supseteq Y$ or X and Y are disjoint.

The first one is a recursive definition. The second one is not. Knuth shows that they are extensionally equivalent. This is a sign that a definition using recursion by name such as (5a) can be avoided if it is not truly necessary.

It may not be necessary, but is it adequate? The answer depends on what we are studying. One striking discovery in mathematics was that recursion by name (reentrancy) can be written without names or labels, using for example paradoxical or fixpoint combinators. If we define the combinators as (6a–b), we get the characteristic equation of recursive behavior in (6c–d).

(6) a. $\mathbf{Y} \stackrel{\text{def}}{=} \lambda h.(\lambda x.h(x\,x))(\lambda x.h(x\,x))$ Curry and Feys (1958)

b. $\mathbf{U} \stackrel{\text{def}}{=} (\lambda x\lambda y.y(xxy))(\lambda x\lambda y.y(xxy))$ Turing (1937)

c. $\mathbf{Y}h = h\,(\mathbf{Y}h) = h\,(h\,(\mathbf{Y}h)) = \cdots$

d. $\mathbf{U}h = h\,(\mathbf{U}h) = h\,(h\,(\mathbf{U}h)) = \cdots$

Notice that neither **Y** nor **U** are recursive definitions, yet they capture recursion by name. (Incidentally, this is the foundation for compiling functional programming

languages, almost all of which are based on lambda calculus. They are all Turing-complete because of this reason.)

The conversion from reentrant (named) recursion to nameless recursion is quite instructive about the powers of recursion by name. Consider the recursive definition of Fibonacci numbers in (7a). It is shown in one piece in (7b), which is then turned into nameless recursion by a series of equivalences in (7c). Notice that h is not recursive by name (f is now a bound variable, which in principle can be eliminated). Its recursion is handled by **Y**.

(7) a. $fib(n) = fib(n-1) + fib(n-2)$ $fib(0) = 0, fib(1) = 1$
b. Let $fib = \lambda n.$if $(n == 0)$ 0 else if $(n == 1)$ 1 else $fib(n-1) + fib(n-2)$
c. Let $h = \lambda f \lambda n.$if $(n == 0)$ 0 else if $(n == 1)$ 1 else $f(n-1) + f(n-2)$
Then $fib = h\, fib$ because $fib\ n = h\, fib\ n,\ \forall n \geq 0$, and $fib = \mathbf{Y}h$ because $fib\ n = \mathbf{Y}h\, n, \forall n,$ and $\mathbf{Y}x = x(\mathbf{Y}x), \forall x$

But this solution comes with a price. **Y** and **U** are not finitely typeable, therefore their solution space cannot be enumerated. Function h is finitely typeable, but *fib* is not just h but $\mathbf{Y}h$, which is not finitely typeable.

This result seems to fly in the face of the fact that *know-, think*-like verbs, and possession-like predicates, are lexical items, therefore they must be finitely typeable and representable. In other words, capturing the meaning of *know* by the formula $\mathbf{Y}know'$ or $\mathbf{U}know'$, where $know'$ is the meaning of *know*, could not stand in for the native speaker's understanding of *know*. Therefore, it is not adequate to use recursion by name in any form to represent competent knowledge of words. Because that kind of knowledge in words is the building block of meaning for syntactic trees, where the meaning of a phrase is combined from the meaning of the parts and the way they are combined, it is not adequate to use recursion by name for syntactic trees of natural strings of words either.

In summary, we have two kinds of evidence that recursion in syntax is not recursion by name, or recursive reentrancy. One is theoretical, as just seen. The other one is empirical, as argued after the example in (4).

7.4 Recursion by Name Is Probably Not Natural

Is recursion by name good for anything? It is indeed. The point is slightly tangential to the purpose of current work, but it allows us to see that in places where recursion by name is necessary and adequate, it is difficult to see a natural phenomenon.

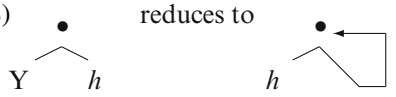
One such domain is programming. With the exception of Fibonacci, the examples we have seen so far are all peripheral recursion, i.e. the recursive value appears on the edge of a tree, which then reiterates, branching on the right edge (2), or the left (3). However, nontail or nonperipheral recursion is possible, and it is not reducible to traversing one periphery of a tree. For example, the pseudo-code below traverses a tree in what is called 'in-order':

```
visit(tree):
  if tree is not nil:
    visit(tree.left)
    print(tree.root)
    visit(tree.right)
end
```

Theory of compiling has shown that we can indeed eliminate such recursion altogether as well, but at the expense of manipulating an auxiliary stack for *every* nonperipheral use of recursion. Such solutions also need elaborate run-time mechanisms to keep track of sequence of computations. It will be clear later that this is fundamentally different than managing an auxiliary stack *once*, per rule, which seems to have natural counterparts.

Thus we either face elimination of recursion by name by fixpoint combinators, which are not finitely typeable, or its elimination by auxiliary devices where every recursive call needs extra stack management. None of these seems to be a natural mechanism.

It is also worth noting the semantics of recursion in programming:

(8) 

We can see the pseudo-code above as a realization of this syntax and semantics, applied twice. Notice the identity of the general mechanism (8) to the structure we considered to be inadequate for natural language recursion, shown in (4). Unlike natural resources such as words, a piece of code can be reentrant; it can point back to its root (and note that the pseudo-code itself is finitely representable), with the understanding that auxiliary mechanisms (such as activation records of recursive calls) and other conventions take care of the rest. None of these mechanisms or their functional equivalent have been attested in cognition, in language, vision, planning, music, or reasoning.

7.5 Recursion by Value Is Not Species-Specific

We can now assume that natural recursion is recursion by value. Given this conception, it is not difficult to see rudiments of recursion in close cousins of ours (if not in other higher animals), especially in planning.

Planning has a long history in AI; see e.g. Ghallab et al. (2004) for an extensive coverage of techniques and tools. Collaborative plans and their relation to psychological states have been extensively studied too; see for example Lochbaum (1998), Bratman (1992), Petrick and Bacchus (2002), Steedman and Petrick (2007), Grosz and Kraus (1993), and Grosz et al. (1999).

The field has devised ingenious ways to capture the act and knowledge of plans as states, search, knowledge representation, and inference. For our purposes, it seems convenient to classify planning in an automata-theoretic way, independent of the aspects above, to highlight its close ties to language.

From this perspective, we can conceive plans at three levels of resource management (with automata-theoretic substrates in parentheses):

(9) a. Reactive planning (finite-state automata—FSA)
b. Instrumental planning (push-down automata—PDA)
c. Collaborative planning (embedded PDA—EPDA)

Finite-state plans deliver whatever organization can be afforded by a finite history and non-embedded behavior. This is not much; for example we cannot capture a scenario where separate actions of an agent match step by step, or a case where a step of the plan needs the result of another plan, either by the same planner or by someone else.

We can model such organized behavior to some extent with PDAs. An example from Jaynes in animal cognition is on the mark (he used it to show deceit as a form of animal consciousness): Jaynes (1976: 219) reports of a chimpanzee in captivity filling his mouth with water in order to penalize a not-so-friendly keeper. The chimpanzee coaxes the keeper, and tries to lure him to proximity to spit water in his face. Sometimes the plan fails, and we would expect the chimpanzee not to spit water. (He might spit water, but not for that purpose. Spit therefore means something more as part of a plan.) His actions depend on how much the keeper conforms to his role as part of the chimpanzee's plan. In this sense it is instrumental planning.

It is worth formalizing some aspects of this planned action to see that at this level of instrumentalizing we are dealing with context-free agent-centered dependencies. Below is a context-free grammar on behalf of the chimpanzee for some potential ways to get what he wants.

(10)

S	→	FillWater LureKeeper Spit
LureKeeper	→	Coax | Hail
Coax	→	Stalk Coax | AskBanana

Spitting depends on achieving LureKeeper, which might enter the chimpanzee's plan by perceiving and interpreting the actions of the keeper. If coaxing him fails, we might still have some acts of stalking the keeper by the chimpanzee, but they would presumably not amount to a plan of spitting at him. This is a dependency that—let's say—he himself established as part of the plan.

From an external observer's point of view the plan-action sequences suggested by this grammar may appear to be finite-state, and indeed this grammar captures a finite-state (regular) language. From the chimpanzee's point of view, it is context-free. I will not extend this way of thinking to suggest that chimpanzees (and bonobos and gorillas) are capable of devising grammars that are strictly context-free from both the observer's and the planner's point of view, but the distinction remains that, unlike reactive planning, an instrumental grammar can be context-free in some way.

This way of thinking coincides with a change of mind in cognitive science. Tomasello and Call (1997) had argued earlier that chimpanzees don't have a mind, but they changed their position in Tomasello et al. (2003), where the finding is that they might have a mind. Crucially, it depends on being aware of other agents, and of potential results that might suggest alternate courses of action if the other agents' actions do not meet the expectations of the concerned chimpanzee from them. This is semantic recursion, or recursion by value, and it is instrumental.

It is sometimes suggested that strict context-freeness and nontrivial recursion can be observed in birdsong as well, in the sense that they sing phrases that seem to consist of subphrases, in one claim to the extent of beyond context-freeness (Stabler 2013). It is not clear to me that we are facing semantic *re*cursion here, because it is not clear that this is not a phonological skill (Berwick et al. 2011, 2013). For it to be semantically compositional the internal phrases must have semantics all the way up, which would indeed be recursion by value. Birdsong might have global semantics, such as happiness, gathering, etc., or make use of very simple rules (Van Heijningen et al. 2009).

7.6 Human Recursion

Availability of other kinds of recursion in humans is not contested by Hauser et al. (2002) and Fitch et al. (2005).[4] They acknowledge that spatial reasoning, among other things, is recursive as well, for example *((((the hole) in the tree) in the glade) by the stream)*, from Fitch et al. (2005). But, once we have a closer look at nonsyntactic recursion, and subsequently at its striking *computational* similarity to language (Sect. 7.6.2), the Chomskyan argument that what makes syntax unique to humans—which I do not dispute—is recursion in it, weakens. A certain kind of recursion may be the most unique human capacity.

As we have seen in Sect. 7.5, instrumental planning can be taken for granted for humans. For a single agent not collaborating with anyone, but perhaps interacting with others, it is easy to see what Searle (1990) called i-intentions, for example scurrying in the park because of rain, to use an example of Searle's. Everyone in the park might target the same shelter, but these would not be we-intentions but a collection of i-intentions. A collection of i-intentions does not constitute a we-intention, Searle claims. This makes perfect sense when we consider dancing in the rain, which might involve the same set of movements as scurrying in the rain from an external observer's point of view, but we know a collaborative dance when we see one, as a collective intention, therefore behavioral equivalence is not the right criterion. Even if a dancer makes a mistake, we would not equate that with an independent act such as someone failing to reach shelter.

[4] See Jackendoff and Pinker (2005) and Parker (2006) for counterarguments on evolutionary basis of syntactic recursion.

The point of collaborative planning and action is that a collection of individuals may intend to pursue a common goal, although they may serve it by different courses of action. In American football, another of Searle's examples, a quarterback and a runningback may have the same intent and execute the same plan, while carrying out different actions.

We can formally incorporate i-intentions, we-intentions, i-plans, and we-plans. Consider the following grammars as proxies for modeling such behavior arising from an internal mechanism. Let us assume extensionally identical (or equivalent) behavior, i.e. dancing and scurrying involve the same actions and are distinguishable only by the intent. The first grammar below is meant for i-intentions and i-plans, and the second one for we-intentions, we-plans, and i-actions.

(11) $S_i \rightarrow \alpha_i \mid A_i$ where α_i is a plan with a base case A_i (Scurry in rain grammar)
$A_1 \rightarrow$ run, and do S_1's work
$\vdots$
$A_n \rightarrow$ run, and do S_n's work

(12) $S_i \rightarrow S'_i[\pi\{S_1, .., S_n\}]$ $\pi\ x$: an ordering of set x (Dance in rain grammar)
$S'_i[S_j \ldots] \rightarrow \alpha_j\, S'_i[\ldots]\, \beta_j$
$\vdots$
$S'_i[S_i] \rightarrow$ run
and do S_i's work A_i

Here is my convention: the grammars are individuated per person i, with the start symbol S_i. We-intentions first make a note of the 'we', using the first rule in (12). In principle it can be of indeterminate number. These rules make use of the Linear-Indexed Grammar (LIG) convention (and the choice is not accidental; cf. subsequent sections). The stack associated with a nonterminal B is denoted $[x \ldots]$ after B, where x is the top. What can enter a stack is planner's decision. Plans and intentions are the righthand sides of rules. Actions and assumptions (or knowledge states) are members of the righthand sides; α and β stand in for contextualizing a participant's action with others, to be differentiated from her own actions, A_i, but related to it structurally, as the unfolding of the mechanism exemplifies for participant k in Fig. 7.1.[5]

[5]I am not suggesting that (12) is the universal schema for all plans. It is meant to show that collaborative plans may be LIG-serializable. LIG-plan space remains to be worked out. For example, base cases of individuated grammars, the $S'_i[S_i]$ rules, doing running—as part of a dance—and A_i would be by definition LIG-serializable too, but in a manner different than what α and β are intended to capture, viz. contextualized knowledge states of the group constituting the we-intention. A_is may be LIG-realized action sequences, making the whole collection a we-plan.

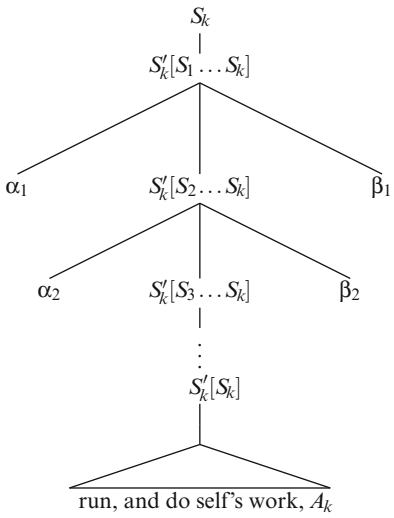


Fig. 7.1 Participant k's grammar for collaborative dancing in the rain

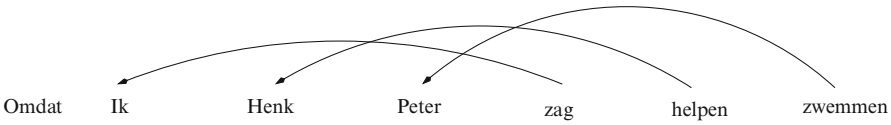
Therefore, every participant can go her own way of carrying out the plan, symbolized by the base cases of her own grammar (the '$S'_i[S_i]$' rules), but not behaving incognizant of the overall plan, symbolized by the top rule, or impervious to other participants' plans and actions, symbolized by the left and right contexts of her own actions/states at the bottom.

If somebody's participation goes wrong, say p's, another participant would know this by observing the failure of S_p for the participant, and recovery may be attempted. (Therefore we assume that the grammar is solving a hidden-variable problem, where it is couched between perception and inference from the world, under the guidance of LIG-automata providing the search space. Zettlemoyer and Collins (2005) started explicit modeling of acquisition of this nature for the case of language.)

No such mechanism is manifest in an instrumental plan such as (11). Thus we can representationally distinguish we-intentions from i-intentions.

7.6.1 Embedded Push-Down Automata for Syntactic Recursion

It has been known since the work of Shieber (1985) that human languages are not context-free. (Most of the earlier proofs turned out to be problematic. Shieber's proof is the one everyone accepts.) The emerging formal characteristics that are adequate for natural languages are summed up in the mild context-sensitivity hypothesis of Joshi (1985): constant growth of string length, i.e. incremental build-up, polynomial parsability, limited cross-serial dependencies, and proper inclusion of context-freeness.



'because I saw Henk help Peter swim ...'

Cross-serial dependencies are the most challenging automata-theoretic aspect to context-freeness. PDAs cannot handle the Dutch dependencies shown above.

However, not all cross-serial dependencies are mildly context-sensitive. Although there are mildly context-sensitive grammars for $\{a^n b^n c^n \mid n \geq 0\}$, which is strictly not context-free, there are no such grammars for $\{www \mid w \in \{a, b, c\}^*\}$, or for $\{w \mid w \in \{a, b, c\}^*$ and $|w|_a = |w|_b = |w|_c\}$.

The last one, called MIX, symbolizes the result that there is probably no human language which is truly scrambling in every word order if mild context-sensitivity is the upper bound, which was an early conjecture of Joshi (1983), which was recently proven by Kanazawa and Salvati (2012). The first one, the "copy" language, shows that human languages do not need queue automata.[6] Together they show that the automata-theoretic approach to linguistic explanation captures some natural boundaries of human syntax and parsing without further stipulation.[7]

Formally speaking, mild context-sensitivity does not define a formal class but makes explicit some desirable and discernible properties. The least powerful extension of context-freeness is a formal class, called linear-indexed languages (Gazdar 1988). Lexicalized tree-adjoining grammars and Combinatory Categorial Grammars (CCG; Steedman 2000) are provably linear-indexed (Joshi et al. 1991).

Linear-indexed languages have an additional property: not only are they polynomially parsable (all MCSLs are), they are efficiently parsable, which means they incur a small polynomial cost in parsing. Dutch and Swiss German data, which exhibit strictly noncontext-free dependencies, can be given a linear-indexed

[6]We note that the language $\{ww \mid w \in \{a, b, c\}^*\}$ is fundamentally different than double-copy $\{www \mid w \in \{a, b, c\}^*\}$. The first one allows stack processing. Here is a LIG grammar for it: $S_{[...]} \rightarrow x\, S_{[x...]}, S_{[...]} \rightarrow S'_{[...]}, S'_{[x...]} \rightarrow S'_{[...]}\, x, S'_{[\,]} \rightarrow \epsilon$, for $x \in \{a, b, c\}$.

[7]Continuing in this way of thinking, we could factor recursion and other dependencies in a grammar, and incorporate word order as a lexically specifiable constraint. It might achieve the welcome result of self-constraining recursion and levels of embedding in parsing: see Joshi (2004: 662).

Both LTAG and CCG avoid recursion by name, LTAG by employing adjunction in addition to substitution, and CCG by avoiding any use of paradoxical combinators such as **Y**, or generalized composition. That is how they stay well below Turing equivalence that might otherwise have been achieved because of recursion by name; see also Joshi (1990), Vijay-Shanker and Weir (1993), and Bozsahin (2012) for discussion of these aspects. Their restrictiveness (to LIG) becomes their explanatory force.

treatment; there are LTAG and CCG grammars for them.[8] Swiss German and Dutch cases can be shown to be abstractly equivalent to $n_1 n_2 n_3 \cdots v_1 v_2 v_3 \cdots$ where n_i is an argument of the verb v_i. In linear-indexed notation we get:

$$
\begin{aligned}
S_{[\ldots]} &\rightarrow n_i \; S_{[i\ldots]} \\
S_{[\ldots]} &\rightarrow S'_{[\ldots]} \\
S'_{[i\ldots]} &\rightarrow \; S'_{[\ldots]} \; v_i \\
S'_{[\,]} &\rightarrow \epsilon
\end{aligned}
$$

The algorithmic substrate of linear-indexed grammars is the Embedded PDA of Vijay-Shanker (1987) and Joshi (1990), which is a stack of stacks (and crucially, not two stacks, a system which we know is Turing-equivalent if they can exchange values). Its grammar formalism passes a single stack among the nonterminals to preserve the dependencies (from left to one symbol on the right, hence the term *linear*). The following grammar is for $\{a^n b^n c^n d^n \mid n \geq 0\}$.[9]

$$
\begin{aligned}
S_{[\ldots]} &\rightarrow a \; S_{[i\ldots]} \; d \\
S_{[\ldots]} &\rightarrow S'_{[\ldots]} \\
S'_{[i\ldots]} &\rightarrow b \; S'_{[\ldots]} \; c \\
S'_{[\,]} &\rightarrow \epsilon
\end{aligned}
$$

7.6.2 *Embedded Push-Down Automata for Human Recursion*

The mechanism that was devised to surpass the context-freeness boundary in syntax is the same as the one we need to move from i-intentions to we-intentions, or from instrumental planning to collaborative multi-agent planning. That seems natural given the relation between organized behavior and serializability, which was first observed in psychology. In Karl Lashley's words:

> Temporal integration is not found exclusively in language; the coordination of leg movements in insects, the song of birds, the control of trotting and pacing in a gaited horse, the rat running the maze, the architect designing a house, and the carpenter sawing a board present a problem of sequences of action which cannot be explained in terms of successions of external stimuli. Lashley (1951: 113)

[8] The Swiss German facts are more direct because the language has overt case marking and more strict word order; see Bozsahin (2012) for a CCG grammar of some Swiss German examples.

[9] Notice that $\{a^n b^n c^n d^n e^n \mid n \geq 0\}$ is not a linear-indexed language, hence such grammars make no use of a linear distance metric, or simple induction from patterns; see Joshi (1983).

Some internal mechanism appears to be at work, and, from the current perspective, LIGs may be the most explicit proposal for the unified problem of characterizing behaviors that are complex enough to rise above data in unexpected ways compared to other kinds of computations by other species.

This is not a resemblance, or reasoning by analogy. It shows that natural languages and natural plans of humans may reduce to the same class of automata-theoretic resource management. If natural computation is what we seek to understand, there seems to be an identifiable mechanism of its management, with many ways to materialize depending on the nature of categories. This does not put humans with language on a par with singing birds and maze-running rats in terms of complexity of organized behavior, but it helps us to understand what added computational explanation is brought in by identifying a class of automata with these behaviors, and the ensuing kinds of recursion that these species are assumed to be capable of.

7.7 Discussion

The preceding argument is not a conjecture that humans must have a general problem solving ability, one omnipotent induction machine without resource boundedness, and that language and planning fall under it. Quite the contrary, there are unique language-specific constraints, much of which have been worked out theoretically. (We cannot say the same thing about we-intentions and collaborative planning. But there is one conjecture of this way of thinking: possible plans may be the LIG-serializable ones.) Languages do not differ arbitrarily, modulo their lexicons. And even there we can expect to see some predictability once we clarify the concept of natural recursion, such as finite representability of lexical items, which in effect rules out any use of recursion by name in the lexicon.

And clearly, language cannot work with action categories, or action with linguistic categories, or music with linguistic categories or with visual ones. Perception is not an omnipotent mechanism. What makes the cognitive processes learnable may be the specialized categories, sort of Humean rise above experience. It does not follow that we learn how to combine in each cognitive domain, rather than combine to learn with some specialized categories.

Deacon (1997) argued that language and brain co-evolved. This proposal bears on the claims for a common substrate. Brain areas that are taken over by language, over at least two million years, are related to planning and action sequencing. Jaynes (1976) had a different agenda, to explain consciousness, which he claimed happened much more recently compared to language, but nevertheless tapping onto the same parts of the brain, and onto the same functionality: combinatory competence. In this regard the practice of writing grammars for language and for planning is not just a historical accident or convention. Grammars are hidden-variables, where the observed form and deduced meaning are hypothesized to be indirectly related by an unobservable grammar.

The automata-theoretic approach to the problem suggests that maybe what we are dealing with is neither deduction nor grammar induction in a naive sense, but computational explanation under resource boundedness constraints (thus making an attempt to avoid the problems of induction, and problems of deduction; cf. Burns 2009; Kok 2013). Out of a possible space of grammars predicted by the identified substrate of automata, it will carve out those which are maximally consistent with perceived data under the constraints of computational complexity. Probably Approximately Correct (PAC) learning is very relevant in this regard: "Inherent algorithmic complexity appears to set serious limits to the range of concepts that can be learned" (Valiant 1984). The class of automata properly identified makes the hypothesis space enumerable for grammars and plans (and for recursion in them), which is one requirement for PAC learning. Finding a hypothesis in polynomial time which is maximally consistent with data is another requirement, and we can maintain P property (for a problem which would be in NP if it is decidable), if we look at likely meanings for words and plans, rather than possible meanings as Quine (1960) did. This is the task of obtaining a grammar by solving a hidden-variable problem, in effect saying that recursion by value is learnable too.

Another purported impediment to PAC learnability of such knowledge is the assumed identity of the data distribution for the sampling of training *and* novel examples, as pointed by Aaronson (2013: 291). However, all that PAC class of learners requires is that the distribution is known, not necessarily *derived* only from early experience. (In this sense, his example of "learning of an *Einstein*" might stray PAC into weird corners of the distribution compared to a mere mortal, which means Nature would have to sample a bit more for him or someone like him to arrive again, but it would sample from the then-current population just like before. In a more recent reassessment of PAC, Valiant (2013) elaborates on Invariance and Learnable Regularity assumptions in relation to natural phenomena such as evolution and mind).

A PAC-like mechanism can safely depend on recursion by value because of its finite representability and its empirical foothold (after all, it is a *value*). The other alternative, reentrancy, or recursion by name as conceived in theoretical computer science, is difficult to assess naturally. There is something unnatural about it. Empirically, it does not correspond to other natural dependencies, which seem to be resource-sensitive and finitely representable. Theoretically, it can be reduced to nameless recursion, which means reduction to recursion by value by a sequence of base cases, which is not enumerable without them.

The last point is equivalent to being uncomputable, for we currently know no way of computing with transfinite representations.[10] We can compute indices of π indefinitely, but we cannot entertain questions regarding the next number after π. Nor can we ask questions about what happens after a computational process fails to halt, and expect an answer.

[10]Notice that lazy evaluation is not a remedy here. By lazy evaluation, we can represent infinite streams by finite means (Abelson et al. 1985; Watt 2004), but for that to work infinite streams must be enumerable.

7.8 Conclusion

From a computer science perspective, natural language syntax does not seem to operate on recursion by name. The kind of dependencies we capture in linguistics when we draw trees of hierarchical structures is recursion by value, which is semantic in nature, but clearly syntacticized. The same can be said about plans, which fall into environment- and object-orientation by affordances (Gibson 1966), syntactically corresponding to type-raising, and to event-orientation by combinatory composition, which is, syntactically, function composition (Steedman 2002).

Recursion-by-value assumption is commonplace in all of cognitive science, also assumed by those who insist it is not needed in syntax (see Everett's commentaries after the recursion conference—Speas and Roeper 2009). I believe that Everett's view is not sustainable (Footnote 1, also Bozsahin 2012), but its failure will not vindicate (Hauser et al. 2002; Fitch et al. 2005).

Humans appear to be uniquely capable of recursion by value, of the kind that can be afforded by a stack of stacks. Various predictions about syntax and other cognitive processes follow from an automata-theoretic way of thinking about them. Therefore, uniqueness of syntax arguments to humans, which I take to be a fact, can be better explained if we conceive automata-constrained recursion as the most unique human capacity for cognitive processes.

Acknowledgements Thanks to PT-AI reviewers and the audience at Oxford, İstanbul, and Ankara, and to Julian Bradfield, Aravind Joshi, Simon Kirby, Vincent Müller, Umut Özge, Geoffrey Pullum, Aaron Sloman, Mark Steedman, and Language Evolution and Computation Research Unit (LEC) at Edinburgh University, for comments and advice. I am to blame for all errors and for not heeding good advice. This research is supported by the GRAMPLUS project granted to Edinburgh University, EU FP7 Grant #249520.

References

Aaronson, S. (2013). Why philosophers should care about computational complexity. In B. J. Copeland, C. J. Posy, & O. Shagrir (Eds.), *Computability: Turing, Gödel, Church, and Beyond*. Cambridge: MIT.

Abelson, H., Sussman, G. J., & Sussman, J. (1985). *Structure and interpretation of computer programs*. Cambridge: MIT.

Berwick, R. C., Okanoya, K., Beckers, G. J., & Bolhuis, J. J. (2011). Songs to syntax: The linguistics of birdsong. *Trends in Cognitive Sciences, 15*(3), 113–121.

Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in Cognitive Sciences, 17*(2), 89–98.

Bozsahin, C. (2012). *Combinatory linguistics*. Berlin/Boston: De Gruyter Mouton.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review, 101*(2), 327–341.

Burns, S. R. (2009). The problem of deduction: Hume's problem expanded. *Dialogue 52*(1), 26–30.

Chomsky, N. (1995). *The minimalist program*. Cambridge: MIT.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry, 36*(1), 1–22.

Chomsky, N. (2013) Problems of projection. *Lingua, 130*, 33–49.

Curry, H. B., & Feys, R. (1958). *Combinatory logic*. Amsterdam: North-Holland.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the human brain*. London: The Penguin Press.
Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology, 46*(4), 621–646.
Everett, D. L. (2009). Pirahã culture and grammar: A response to some criticisms. *Language, 85*(2), 405–442.
Fitch, T., Hauser, M., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition, 97*, 179–210.
Gazdar, G. (1988). Applicability of indexed grammars to natural languages. In U. Reyle & C. Rohrer (Eds.), *Natural language parsing and linguistic theories* (pp. 69–94). Dordrecht: Reidel.
Ghallab, M., Nau, D., & Traverso, P. (2004) *Automated planning: Theory and practice*. San Francisco: Morgan Kaufmann.
Gibson, J. (1966). *The senses considered as perceptual systems*. Boston: Houghton-Mifflin Co.
Grimshaw, J. (1990). *Argument structure*. Cambridge: MIT.
Grosz, B., & Kraus, S. (1993). Collaborative plans for group activities. In *IJCAI*, Chambéry (Vol. 93, pp. 367–373).
Grosz, B. J., Hunsberger, L., & Kraus, S. (1999). Planning and acting together. *AI Magazine, 20*(4), 23.
Hale, K., & Keyser, S. J. (2002). *Prolegomenon to a theory of argument structure*. Cambridge: MIT.
Hauser, M., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569–1579.
Huybregts, R., & van Riemsdijk, H. (1982). *Noam Chomsky on the generative enterprise*. Dordrecht: Foris.
Jackendoff, R., & Pinker, S. (2005). The nature of the language faculty and its implications for language evolution. *Cognition, 97*, 211–225.
Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. New York: Houghton Mifflin Harcourt.
Joshi, A. K. (1983). Factoring recursion and dependencies: An aspect of tree adjoining grammars (TAG) and a comparison of some formal properties of TAGs, GPSGs, PLGs, and LPGs. In *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, Cambridge (pp. 7–15)
Joshi, A. (1985). How much context-sensitivity is necessary for characterizing complex structural descriptions—Tree adjoining grammars. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 206–250). Cambridge: Cambridge University Press.
Joshi, A. (1990). Processing crossed and nested dependencies: An automaton perspective on the psycholinguistic results. *Language and Cognitive Processes, 5*, 1–27.
Joshi, A. K. (2004). Starting with complex primitives pays off: Complicate locally, simplify globally. *Cognitive Science, 28*(5), 637–668.
Joshi, A., & Schabes, Y. (1992). Tree-adjoining grammars and lexicalized grammars. In M. Nivat & A. Podelski (Eds.), *Definability and recognizability of sets of trees*. Princeton: Elsevier.
Joshi, A., Vijay-Shanker, K., & Weir, D. (1991). The convergence of mildly context-sensitive formalisms. In P. Sells, S. Shieber, & T. Wasow (Eds.), *Foundational issues in natural language processing* (pp. 31–81). Cambridge: MIT.
Kanazawa, M., & Salvati, S. (2012). MIX is not a tree-adjoining language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, Jeju Island (pp. 666–674). Association for Computational Linguistics.
Knuth, D. E. (1968). *Fundamental algorithms* (The art of computer programming, Vol. 1). Reading: Addison-Wesley.
Kok, A. (2013). Kant, Hegel, und die Frage der Metaphysik: Über die Möglichkeit der Philosophie nach der Kopernikanischen Wende. Wilhelm Fink.
Lashley, K. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley. Reprinted in Saporta (1961).

Lobina, D. J., & García-Albea, J. E. (2009). Recursion and cognitive science: Data structures and mechanisms. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1347–1352).

Lochbaum, K. E. (1998). A collaborative planning model of intentional structure. *Computational Linguistics, 24*(4), 525–572.

Manning, C. D. (1996). *Ergativity: Argument structure and grammatical relations*. Stanford: CSLI.

Nevins, A., Pesetsky, D., & Rodrigues, C. (2009). Pirahã exceptionality: A reassessment. *Language, 85*(2), 355–404.

Parker, A. R. (2006). Evolving the narrow language faculty: Was recursion the pivotal step. In *The Evolution of Language: Proceedings of the 6th International Conference on the Evolution of Language* (pp. 239–246). Singapore: World Scientific Press.

Petrick, R. P., & Bacchus, F. (2002). A knowledge-based approach to planning with incomplete information and sensing. In *AIPS*, Toulouse (pp. 212–222).

Peyton Jones, S. L. (1987). *The implementation of functional programing languages*. New York: Prentice-Hall.

Quine, W. v. O. (1960). *Word and object*. Cambridge: MIT.

Saporta, S. (Ed.). (1961). *Psycholinguistics: A book of readings*. New York: Holt Rinehart Winston.

Searle, J. R. (1990). Collective intentions and actions. In P. R. Cohen, M. E. Pollack, & J. L. Morgan (Ed.), *Intentions in communication*. Cambridge: MIT.

Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy, 8*, 333–343.

Speas, M., & Roeper, T. (Eds.). (2009, forthcoming). *Proceedings of the Conference on Recursion: Structural Complexity in Language and Cognition*, University of Mass, Amherst.

Stabler, E. (2013). *Copying in mildly context sensitive grammar*. Informatics Seminars, Institute for Language, Cognition and Computation, University of Edinburgh, October 2013.

Steedman, M. (2000). *The syntactic process*. Cambridge: MIT.

Steedman, M. (2002). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy, 25*, 723–753.

Steedman, M., & Petrick, R. P. (2007). Planning dialog actions. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue (SIGdial 2007)*, Antwerp (pp. 265–272)

Tomasello, M., & Call, J. (1997). *Primate cognition*. New York: Oxford University Press.

Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states—the question is which ones and to what extent. *Trends in Cognitive Sciences, 7*(4), 153–156.

Turing, A. M. (1937). Computability and λ-definability. *Journal of Symbolic Logic, 2*(4), 153–163.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. New York: Basic Books.

Van Heijningen, C. A., De Visser, J., Zuidema, W., & Ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proceedings of the National Academy of Sciences, 106*(48), 20538–20543.

Vijay-Shanker, K. (1987). *A study of tree adjoining grammars*. PhD thesis, University of Pennsylvania.

Vijay-Shanker, K., & Weir, D. (1993). Parsing some constrained grammar formalisms. *Computational Linguistics, 19*, 591–636.

Watt, D. A. (2004). *Programming language design concepts*. Chicester: Wiley.

Zettlemoyer, L., & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, Edinburgh.

Part II
Information

Chapter 8
AI, Quantum Information, and External Semantic Realism: Searle's Observer-Relativity and Chinese Room, Revisited

Yoshihiro Maruyama

Abstract In philosophy of mind, Searle contrived two arguments on the impossibility of AI: the Chinese room argument and the one based upon the observer-relativity of computation. The aim of the present article is two-fold: on the one hand, I aim at elucidating implications of the observer-relativity argument to (ontic) pancomputationalism, in particular the quantum informational view of the universe as advocated by Deutsch and Lloyd; on the other, I aim at shedding new light on the Chinese room argument and the nature of linguistic understanding in view of the semantic realism debate in philosophy of logic and language, especially Dummett's verificationist theory of meaning. In doing so, philosophy of mind turns out to be tightly intertwined with philosophy of logic and language: intelligence is presumably the capacity to reason, and in view of a distinction between statistical and symbolic AI ("AI of sensibility" and "AI of understanding" in Kantian terms), philosophy of logic and language is arguably the part of philosophy of mind that concerns the symbolic realm of intelligence (i.e., the realm of understanding rather than sensibility). More specifically, in the first part of the article, I argue that pancomputationalism cannot be maintained under Searle's external realism; nevertheless, a radical (external) antirealist position, such as Wheeler's ("It from Bit"), may allow for a possibility of pancomputationalism. The Searle's argument and the infinite regress paradox of simulating the universe yield challenges to pancomputationalism and the quantum informational view of the universe, leading us to the concept of weak and strong information physics (just like weak and strong AI). In the second part, I argue that Dummett's principle of manifestation on linguistic understanding commits Searle to semantic realism due to the nature of his Chinese room argument. Searle's position must thus be realism in two senses, that is, it has to be external semantic realism. I finally focus upon recent developments of categorical quantum mechanics, and discuss a quantum version of the Chinese room argument. Underpinning all this is the conceptual view that the duality of meaning manifests in different philosophies of logic, language, and mind.

Y. Maruyama (✉)
Quantum Group, Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford, OX1 3QD, UK
e-mail: maruyama@cs.ox.ac.uk

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_8

Keywords Chinese room • Quantum mechanics • John Searle • Pancomputationalism • Michael Wheeler • Quantum pancomputationalism • External semantic realism • Chinese room argument • Observer-relativity of computation • Strong and weak information physics

8.1 Introduction: Searle's Chinese Room and Observer-Relativity Arguments

The Chinese room argument by John Searle is prominent among arguments against the concept of genuine artificial intelligence (such as strong AI or the Turing test), having been intensively discussed by both proponents and opponents (see, e.g., Cole 2009). Less known is his later argument based upon the observer-relativity of computation (Searle 2002), which shall be called the observer-relativity argument in the present article. It basically proceeds as follows.

1. Computation exists relative to observers.
2. However, human intelligence does not.
3. Therefore, the latter cannot be reduced to the former.

For the moment let us put it aside to explicate why computation is relative to observers (though Searle asserts that it is obvious in the quotation below). Interestingly, Searle (2002) concludes the article with the following retrospective remarks (p. 17):

> Computation exists only relative to some agent or observer who imposes a computational interpretation on some phenomenon. This is an obvious point. I should have seen it ten years ago, but I did not.

Although there are still quite some on-going debates on the plausibility of the observer-relativity argument, in the present article, I focus on implications rather than the pros and cons of the argument.

More specifically, the aim of the article is to show that the observer-relativity argument sheds new light upon ontic pancomputationalism, according to which the universe itself is a computing system; especially we focus upon the quantum version of ontic pancomputationalism, namely the view that the universe is a huge quantum computer (we omit "ontic" in the following; see Piccinini (2010) for varieties of pancomputationalism). This sort of quantum informational view of the universe has been advocated by Lloyd (2006), Deutsch (1997), and others including both physicists and philosophers.

The quantum informational view of the universe may appear to be an issue totally different from Searle's philosophy of mind, but it is closely related indeed. Among other things, the observer-relativity of computation seems most obvious in the case of quantum computation, which allows us to exploit microscopic quantum phenomena in order to compute faster and communicate securer than possible in the conventional classical framework. Then, it is we observers that regard the unitary time evolution of a quantum system as a computational process; the former, by itself, is merely a physical phenomenon.

In this direction, I finally argue that a quantised version of the observer-relativity argument refutes a strong form of quantum pancomputationalism as long as the universe is not observer-relative (yet a modest form of it remains maintainable even in that case). To put it the other way around, if we are happy to consider the universe to be observer-relative just as Wheeler (1990) indeed does with the famous saying "It from Bit", then we can still keep the strong quantum informational view consistent. In order to endorse the strong pancomputationalism thesis, we must thus choose either the Searle's realist view or an antirealist view such as Wheeler's. In this way, pancomputationalism is tightly intertwined with the realism/antirealism debate.

How does the observer-relativity argument relate to the Chinese room argument? Just before the remarks above, Searle summarises the Chinese room argument as follows (Searle 2002, p. 17):

> The Chinese room argument showed semantics is not intrinsic to syntax.

To put it differently, syntax is not enough to confer meaning on symbols, or it is not "sufficient for semantic content" in Searle's words. In contrast to this, the point of the observer-relativity argument is summarised as follows (*ibid.*, p. 17):

> But what this argument shows is that syntax is not intrinsic to physics.

The observer-relativity argument is more fundamental than the Chinese room argument in the sense that even if syntax is sufficient for semantics, any computer, which itself is a physical entity, cannot even represent syntax of language because physics alone is not sufficient for syntax. In other words, computation (as syntactic symbol manipulation) is more than mere physics, and thus the computer *per se* does not compute. Computation is only enabled in the presence of both a suitable physical system and an observer regarding the time evolution of the system as a computational process. From the Searle's point of view, therefore, computation is necessarily human computation as it were; there is no computation whatsoever in the absence of observers (yet it is not clear whether non-human beings can count as observers in Searle's view).

Now, let us turn to implications of the Chinese room argument, the second topic of the article. Searle (2002) asserts that syntax by itself is not sufficient for semantic content (p. 16):

> In all of the attacks on the Chinese room argument, I have never seen anyone come out baldly and say they think that syntax is sufficient for semantic content.

Would it really be impossible to account for semantics in terms of syntax? What is called proof-theoretic semantics (see, e.g., Kahle and Schröder-Heister 2005) may be seen as a sort of way to do it. Proof-theoretic semantics is an enterprise to account for the meaning of logical and other expressions in terms of proof theory within the tradition of Gentzen, Prawitz, and Martin-Löf. It has a philosophical origin in Dummett's antirealist philosophy, and may be regarded as a form of inferentialism as advocated by Brandom (2000). Traditionally, the enterprise of semantics was mostly along the line of the referentialist or denotationalist account of meaning,

such as the Tarski semantics, in which to understand a sentence is to know its truth-conditions through the denotations of expressions involved. It is still the dominating paradigm of semantics in many fields of pure and applied logic.

Proof-theoretic semantics objects to it, claiming that the meaning of a word can fully be given by the inferential role it plays in our linguistic practice, without any reference to objects outside language. Some proponents of proof-theoretic semantics refer to the later Wittgenstein's thesis "Meaning is use." (In light of his later philosophy, however, Wittgenstein himself would not think there is any explicit formal rule governing the use of language; this is obviously relevant to the issue of rule following and to the Kripkenstein paradox.) Especially, the meaning of a logical constant is accounted for by the inferential rules governing it (e.g., the introduction and/or elimination rules in the system of natural deduction). Thus, syntax is autonomous and meaning-conferring in proof-theoretic semantics, and we do not need truth conditions or denotations to confer meaning on logical and other symbols. There is no outside syntax in proof-theoretic semantics, and syntax is indeed sufficient for semantics.

A philosophical underpinning of proof-theoretic semantics is Dummett's arguments against semantic realism; especially, in this article, we focus on his manifestation argument, which is based on the principle of manifestation on linguistic understanding. I argue that Searle's conception of linguistic understanding violates the principle of manifestation, and thus he must be committed to semantic realism. While Searle takes the position of external realism on the nature of the universe, his position must be semantic realism as well, which is realism on the nature of meaning or linguistic understanding.

8.2 Observer-Relativity and Pancomputationalism: Keep External Realism or Allow Antirealism?

In this section, we first briefly review the Searle's idea of observer-relativity, and then address implications of the observer-relativity argument to the quantum informational view of the universe, finally leading to the conclusion that quantum pancomputationalism is not tenable as long as a form of scientific realism is maintained in the sense that the universe exists independently of observers; yet antirealism such as Wheeler's allows for a possibility of quantum pancomputationalism. Searle's external realism plays a crucial rôle in the justification of the quantum observer-relativity argument presented below.

Searle (2002) argues in favour of the observer-relativity of computation in the following way (p. 17):

1. Computation is defined in terms of symbol manipulation.
2. The notion of a symbol is not a notion of physics, but a notion of observers (who decide upon whether to regard physical tokens in Nature as symbolic entities).
3. Therefore, computation is not intrinsic to physics, and relative to observers.

There are merely some electromagnetic phenomena going on inside a computer (cf. Landauer (1991)'s dictum "Information is physical"), and the physical phenomena themselves are not computation. The computer is a system of physical devices. Any physical entity *per se* cannot be a symbol, and so cannot constitute syntax consisting of symbols, much less semantics. In a nutshell, the computer *per se* does not compute. Rather, we observers conceive of the physical phenomena as computational processes, and of the computer as computing. Whereas physical phenomena without observers are nothing more than physics, those with observers can be computation. In such a way, we may lead to the Searle's idea that computation is relative to observers. From the Searle's point of view, computation is not a matter of reality, but a matter of observation.

Searle's view may, of course, be contested from different perspectives (one could even argue that, just as computation is relative to observers, intelligence is relative to observers, and so the human does not think just as the computer does not compute); in this section, however, I aim at elucidating what insights can be derived from it, especially in relation to the quantum information view of the universe.

8.2.1 Is Ontic Pancomputationalism Tenable or Not?

Quantum computation is a relatively new, but recently rapidly growing paradigm of models of computation, facilitating much faster and securer ways of computing and communicating than classical computation. There are some other novel models of computation as well. While quantum computation builds upon microscopic physical systems, for example, DNA computation is based on biological systems, utilising their salient features as resources for computation. Searle (2002) succinctly pins down the core idea of such emergent models of computation, in saying "you can assign a computational interpretation to anything" (which is part of the quotation above), even though he does not explicitly touch upon such recent models of computation.

The basic idea of quantum computation (especially, the quantum circuit model) is that quantum states can be seen (by us observers) as information (called qubits), and the unitary time evolution (and measurements) of them as information processing. In a nutshell, we may view quantum dynamics as computational processes, and then we are able to exploit salient features of quantum physics, such as entanglement (or the Einstein-Podolsky-Rosen "paradox"), as resources for computation; it is widely believed in the quantum information community that this way of thinking played a significant role in contriving quantum protocols (e.g., quantum teleportation and superdense coding). Likewise, interpreting DNA dynamics as computational processes leads us to DNA computation. Thus, we observers are always allowed to (and not to) interpret phenomena as computation. In the light of this, I would say that Searle's observer-relativity perspective on computation is not only conceptually important, but also practically matter, indeed lying at the heart of different sorts of so-called natural computing as mentioned above.

At the same time, however, Searle's observer-relativity argument, I think, allows us to make a critical objection to the quantum informational view of the universe. I especially have in mind the claim of Lloyd (2006) that the universe is a quantum computing system. It is similar to the assertion that Nature is computational intelligence. Searle (2002) says:

> The natural sciences describe features of reality that are intrinsic to the world as it exists independently of any observers.

On the other hand, computation is observer-relative, and does not describe intrinsic features according to him. In the light of this, we may adapt the observer-relativity argument presented above to contrive the following, quantum observer-relativity argument:

1. Quantum computation exists relative to observers.
2. However, the universe exists independently of observers.
3. Therefore, the latter cannot be reduced to the former, so that the universe cannot be a quantum computer.

Actually, we do not really have to focus upon quantum computation alone, but rather we may address the possibility of pancomputationalism in general. Nevertheless, there are two reasons not to do so: firstly, non-quantum pancomputationalism is not plausible any more in the light of the quantum nature of the world; secondly, the claim of item 1 is more convincing in the case of quantum rather than classical computation as already noted above (who thought of quantum systems as computing before the discovery of quantum computation? Any quantum system would not have been computing in that classical era).

Obviously, the quantum observer-relativity argument hinges upon the claim of item 2, a form of scientific realism. Accordingly, we may seek a possibility of the quantum information view in the absence of this sort of realism. At the same time, however, Searle himself takes the position of the so-called "external realism", asserting as follows:

> There exists a real world that is totally independent of human beings and of what they think or say about it. (Searle 1998, p. 13)

> There is a way that things are independently of our representations (Searle 1998, p. 31)

We may thus conclude that the Searle's position adopting external realism together with observer-relativity is inconsistent with the quantum informational view of the universe. As already discussed, quantum computation is in good harmony with the observer-relativity thesis, and therefore the only remaining option to maintain the quantum informational view would be to revise external realism in some way or other.

Leaving this issue in the next subsection, It should be noted here that this quantum version of the observer-relativity argument never refutes the possibility of quantum computation qua technology, and does not give any objection to the so-called information (or digital) physics enterprise qua science (it seem so interesting and promising that I am indeed working on it). But rather the point is that even if it

finally succeeds in accounting for the complex physics of the entire universe, it does not *ipso facto* imply that the universe *per se* is a quantum computer, or quantum computational processes.

8.2.2 Strong vs. Weak Theses of Information Physics

Information physics (aka. digital physics with a little bit different meaning) has already gained quite some successes (e.g., the well-knwon informational account of the Maxwell's demon; the operational reconstruction of quantum theory in Chiribella et al. 2011), and it would deserve more philosophical attention. Information physics ultimately aims at reconstructing and developing the whole physics in terms of information, which is taken to be a primary entity, considered to be more fundamental than physical objects. In information physics, it is not that there are computational processes because there are physical systems to implement them, but that there are computational processes in the first place, and physics is just derived from them. Philosophically, this may count as a sort of process philosophy as advocated by Whitehead and Leibniz (under a certain interpretation).

AI and information physics are quite different issues with no apparent link between them (except the concept of computation giving their underpinnings). As already seen in the previous subsection, it seems fruitful to borrow concepts in AI, or philosophy of mind, in order to shed new light on information physics. Just like the common concept of weak and strong AI, I propose to make a distinction between weak and strong IP (Information Physics), or weak pancomputationalism and strong pancomputationalism:

1. Weak IP (weak pancomputationalism) is the view that (some constituents of) the universe may be interpreted as computational processes.
2. Strong IP (strong pancomputationalism) is the view that the universe *per se* is a bunch of computational processes as a matter of fact.

The quantum observer-relativity argument presented above surely refutes the latter, but not really the former. Because interpretation is a matter of observers, and the weak IP view does not hold that the universe is computational processes independently of us observers.

We may conceive of another strong IP view that the universe can be simulated by a computer; Lloyd (2006) alleges it would, in principle, be possible. I think, however, that the notion of the computer simulating the universe would suffer from a logical paradox because it involves the following self-referential infinite regress:

1. The computer simulating the entire universe must simulate itself.
2. This implies that the computer must simulate the computer simulating the universe.
3. Likewise, it must simulate the computer simulating the computer simulating the universe.

4. This continues *ad infinitum*; hence no computer simulating the universe.

This may be called the paradox of simulating the universe (cf. the supertask paradox; the paradox of the set-theoretical universe containing itself as a set).

John Wheeler's dictum "It from Bit" is along a similar line: Wheeler (1990) endorses the following doctrine:

> All things physical are information-theoretic in origin.

However, he boldly thinks that the universe and every "it" in the universe arise from our observations (Wheeler 1990). He thus seems to reject the assertion of item 2 above, maintaining that the universe is actually relative to observers. This might make sense in quantum physics in particular. The relationships between systems and observers are quite subtle in quantum physics: there is no neutral way to see (or measure) quantum systems as they are, without disturbing them through observations, and it is impossible to assign values to all observables (physical quantities) in a coherent way (the Kochen-Specker theorem). Hence we cannot really access the "reality" of quantum systems, which, some people think, do not actually exist; at least, we cannot maintain local realism according to the Bell's theorem. Consequently, it seems plausible to some extent to think that the universe is relative to observers due to its quantum nature.

At the same time, however, Wheeler does not restrict his claim into the quantum realm, applying the antirealist view to classical macroscopic systems as well as quantum microscopic ones. Therefore, it is indeed a radical antirealist position, to which Searle's observer-relativity argument does not apply, and which may be coherent as a philosophical standpoint, at the cost of giving up the ordinary realist view of Nature.

8.3 The Chinese Room and Semantic Realism: What Does the Understanding of Language and Meaning Consist in?

In this section we first have a look at the issue of how external realism relates to semantic realism, and then elucidate Searle's position about the nature of linguistic understanding in view of the semantic realism debate concerning in particular Dummett's philosophy of language. As a case study we also discuss a quantum version of the Chinese room argument in the context of categorical quantum mechanics and quantum linguistics.

As mentioned in the last section, Searle's position on the realism debate is characterised as external realism, which is basically a position on the nature of physical reality or the universe, and as such has nothing to do with the nature of language and meaning, or the nature of linguistic understanding. Semantic realism discussed in this section is primarily about the latter, even though Dummett attempts to relate them by the so-called constitution thesis that "the literal content of realism consists in the content of semantic realism" (Miller 2010); according to Dummett, "the

theory of meaning underlies metaphysics" (just as the denotationalist/verificationist conception of meaning underlies realism/antirealism). By contrast, Devitt (1991) argues (p. 39):

> Realism says nothing semantic at all beyond . . . making the negative point that our semantic capacities do not constitute the world

In such a view, semantic realism and external realism, in principle, have nothing to do with each other. Even so, however, I am going to argue in this section, however, I argue that Searle must commit himself not only to external realism but also to semantic realism, due to his view on the understanding of meaning.

8.3.1 Dummett's Manifestation Argument Leads Searle to Semantic Realism

In philosophy of logic, Dummett's view on the meaning of logical constants has led to what is now called proof-theoretic semantics, as opposed to model-theoretic semantics (in logic, semantics traditionally meant the latter only). From the perspective of proof-theoretic semantics, meaning is inherent in syntactic rules governing how to use symbols, and thus grasping meaning is nothing more than grasping those rules; there is no need of any further elements like truth conditions or denotations. Proponents of proof-theoretic semantics thus consider syntax to be sufficient to confer meaning upon symbols, and so for semantics.

This is the view of proof-theoretic semantics, obviously being in striking contrast with Searle's Chinese room view that syntax alone is not enough to account for semantic content. A philosophical underpinning of proof-theoretic semantics is Dummett's arguments against semantic realism; another is Wittgenstein's thesis "Meaning is use." Dummett (1978), *inter alia*, contrives the so-called manifestation argument, part of which we focus on here.

In this article, semantic realism is characterised as the position that admits "recognition-transcendent" (Dummett's term) contents in the understanding of language. Dummett's puts emphasis on the recognition-transcendency of truth conditions; here, not only truth conditions but also any sort of recognition-transcendent contents are allowed.

Searle (1992) explains the point of the Chinese room argument as follows (p. 45):

> I believe the best-known argument against strong AI was my Chinese room argument that showed a system could instantiate a program so as to give a perfect simulation of some human cognitive capacity, such as the capacity to understand Chinese, even though the system had no understanding of Chinese whatever.

Searle thus thinks that any syntactical or computational ability to simulate language does not, by itself, guarantee the semantic understanding of language. This is the reason why Searle says the Chinese room argument showed that semantics is not intrinsic to syntax.

What is crucial here is the following: it is not that there are some problems on the simulation of language, but that the simulation is perfect, yet it is not sufficient for the understanding of language. Searle indeed uses the term "perfect simulation" in the quotation above. According to him, understanding is more than perfect simulation.

Dummett's manifestation argument against semantic realism is based on the principle of manifestation, which Miller (2010) formulates as follows:

> If speakers possess a piece of knowledge which is constitutive of linguistic understanding, then that knowledge should be *manifested* in speakers' use of the language i.e. in their exercise of the practical abilities which constitute linguistic understanding.

That is, there is no hidden understanding beyond practical capacities to use language in various situations, namely beyond the capacity to simulate language. On the ground that anything manifested in linguistic practice can be simulated, we may conclude that Searle's idea that even perfect simulation is not sufficient for the understanding of language violates the principle of manifestation. And thus Searle is compelled to commit himself to semantic realism.

To put it differently, the principle of manifestation says that the understanding of language must be simulatable; this is Dummett's view. On the other hand, Searle is directly against such a conception of linguistic understanding as seen in his above remarks on the Chinese room argument. We may thus say that Dummett's antirealist view on linguistic understanding is in sharp conflict with Searle's realist view, especially in terms of the manifestability of understanding.

8.3.2 Categorical Quantum Mechanics and Linguistics: Can Quantum Picturalism Confer Understanding of Meaning?

Nearly a decade ago, categorical quantum mechanics (see, e.g., Abramsky and Coecke 2008) paved the way for a novel, high-level (in the technical sense), category-theoretical formalism to express quantum mechanics and quantum computation, thus allowing us to reason about quantum systems via its graphical language and thereby to verify quantum communication protocols and algorithms in a fairly intuitive fashion, with the flows of information exhibited clearly in the graphical language of quantum picturalism (Coecke 2010).

The graphical language of categorical quantum mechanics enables us to dispense with complicated algebraic calculations in the Hilbert space formalism of quantum mechanics, replacing them by simpler graphical equivalences. Still, what is provable is the same, and indeed there is a sort of completeness theorem between categorical quantum mechanics and the standard Hilbert space formalism, which ensures the equivalence between them.

The paper "Kindergarten Quantum Mechanics" (Coecke 2005) claims that even kindergarten students can understand the pictorial language of categorical quantum mechanics, and so quantum mechanics itself. It is just a simple manipulation of pictures consisting of strings, boxes, and so on; thus, children could understand it as Coecke (2005) says. The question is then the following: do those children or computers that are able to manipulate pictures in a suitable way understand quantum mechanics? For example, the quantum teleportation protocol can be verified just by "yanking" in the pictorial language, and then, do such children or computers understand the teleportation protocol? We can get even closer to the original Chinese room argument in the case of quantum linguistics.

Quantum linguistics emerged from the spirit of categorical quantum mechanics, integrating Lambek pregroup grammar, which is qualitative, and the vector space model of meaning, which is quantitative, into the one concept via the methods of category theory. It has already achieved, as well as conceptual lucidity, experimental successes in automated synonymity-related judgement tasks (such as disambiguation). It is equipped with a graphical language in the same style as categorical quantum mechanics. Then, do computers capable of manipulating pictures in quantum linguistics understand language (if quantum linguistics perfectly simulates language)? This is almost the same as the main point of the Chinese room argument. A similar question was raised by Bishop et al. (2013).

In order to address the question, I would like to make a distinction between mathematical meaning and physical meaning. Then, the question turns out to consist of two different questions: if one understands the graphical language of quantum mechanics, then does the person understand the mathematical meaning of quantum mechanics?; and how about the physical meaning?

Here let us assume that the capacity to manipulate symbols (including figures) is sufficient for mathematical understanding. Thus, for example, the mechanical theorem prover does understand mathematics. Under this assumption, the first question may be given an affirmative answer, yet the answer to the second one on physical meaning would be negative. Physical understanding must connect the mathematical formalism with elements of Nature so that the former correctly models the latter. This modelling capacity is more than the mathematical ability to manipulate symbols. Broadly speaking, physical understanding is mathematical understanding plus modelling understanding.

At the same time, however, Searle himself would probably object to the very assumption, since he puts strong emphasis on intensionality. For him, any sort of understanding, including mathematical understanding, could not be gained via the mere capacity to manipulate symbols. He would thus think that the theorem prover does not understand mathematics, even if it can prove more theorems than ordinary mathematicians; he might call it the "mathematical room" argument.

8.4 Concluding Remarks

In the present article, we have revisited the Searle's two arguments, the Chinese room and observer-relativity arguments, in relation to quantum pancomputationalism and the realism debate.

In the first paper of the article, I have argued that quantum pancomputationalism is inconsistent with external realism, yet pancomputationalism is consistent with antirealist positions, which are more or less philosophically demanding, though. To be precise, this is about pancomputationalism in the sense of strong IP; the weak IP view is consistent with pancomputationalism even in the presence of external realism. I also touched upon the paradox of simulating the universe, which is another challenge to pancomputationalism.

In the second part, I have argued that Searle must commit himself not only to external realism, but also to semantic realism, because of his position on linguistic understanding as seen in the Chinese room argument. The argument was based on Dummett's principle of manifestation on linguistic understanding. Finally, I discussed the Chinese room argument in the context of categorical quantum mechanics and its graphical language. We could separate mathematical understanding and physical understanding, and argue that the capacity to manipulate graphical rules is sufficient for mathematical understanding, but not for physical understanding. Searle would think, however, that it is insufficient for both, due to his semantic realism. No theorem prover could understand mathematics from Searle's realist point of view.

Overall, the present article may be regarded as pursuing the duality of meaning in its different guises: the dualities between the model-theoretic/referentialist/realist and proof-theoretic/inferentialist/antirealist conceptions of meaning in philosophy of logic/language/mind. These exhibit duality even in the sense that, whereas referential realism makes ontology straightforward and epistemology complicated (e.g., how to get an epistemic access to independent reality could be a critical problem as exemplified by Benacerraf's dilemma), inferential antirealism makes epistemology straightforward and ontology complicated (e.g., anything apparently existing has to be translated into something else with an equivalent function). Put in a broader context, these dualities could presumably be compared with more general dichotomies between substance-based and function/relation/process-based metaphysics.

References

Abramsky, S., & Coecke, B. (2008). Categorical quantum mechanics. In K. Engesser, D. M. Gabbay, & D. Lehmann (Eds.), *Handbook of quantum logic and quantum structures* (pp. 261–324). Amsterdam: Elsevier.

Bishop, J. M., Nasuto, S. J., & Coecke, B. (2013). 'Quantum linguistics' and Searle's Chinese room argument. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence* (pp. 17–28). Berlin: Springer.

Brandom, R. (2000). *Articulating reasons: An introduction to inferentialism*. Cambridge: Harvard University Press.

Chiribella, G., D'Ariano, G. M., & Perinotti, P. (2011). Informational derivation of quantum theory. *Physical Review A, 84*, 012311.

Coecke, B. (2005). Kindergarten quantum mechanics. *AIP Conference Proceedings, 810*, 81–98. American Institute of Physics.

Coecke, B. (2010). Quantum picturalism. *Contemporary Physics, 51*, 59–83.

Cole, D. (2009). The Chinese room argument. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

Dennett, D. (1978). Toward a cognitive theory of consciousness. In D. C. Dennett (Ed.), *Brainstorms: Philosophical essays on mind and psychology* (pp. 149–173). Cambridge: MIT.

Deutsch, D. (1997). *The fabric of reality*. London: Penguin.

Devitt, M. (1991). *Realism and truth* (2nd ed.). Princeton: Princeton University Press.

Dummett, M. (1978). *Truth and other enigmas*. London: Duckworth.

Kahle, R., & Schröder-Heister, P. (Eds.). (2005). *Proof-theoretic semantics* (Special Issue of Synthese, Vol. 148). Berlin: Springer.

Landauer, R. (1991). Information is physical. *Physics Today, 44*, 23–29.

Lloyd, S. (2006). *Programming the universe: A quantum computer scientist takes on the cosmos*. New York: Knopf.

Miller, A. (2010). Realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

Piccinini, G. (2010). Computation in physical systems. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge: MIT.

Searle, J. (1998). *Mind, language, and society: Philosophy in the real world*. New York: Basic Books.

Searle, J. (2002). The problem of consciousness. In J. R. Searle (Ed.), *Consciousness and language* (pp. 7–17). Cambridge: Cambridge University Press.

Wheeler, J. A. (1990). Information, physics, quantum: The search for links. In W. H. Zurek (Ed.), *Complexity, entropy, and the physics of information* (pp. 309–336). Redwood City: Addison-Wesley.

Chapter 9
Semantic Information and Artificial Intelligence

Anderson Beraldo de Araújo

Abstract For a computational system to be intelligent, it should be able to perform, at least, basic deductions. Nonetheless, since deductions are, in some sense, equivalent to tautologies, it seems that they do not provide new information. In order to analyze this problem, the present article proposes a measure of the degree of semantic informativity of valid deductions. Concepts of coherency and relevancy, displayed in terms of insertions and deletions on databases, are used to define semantic informativity. In this way, the article shows that a solution to the problem about informativity of deductions provides a heuristic principle to improve the deductive power of computational systems.

Keywords Semantic information • Artifial intelligence • Scandal of deduction

9.1 Introduction

For Aristotle, "every belief comes either through syllogism or from induction" (Aristotle 1989). From that, we can infer that every computational system that aspires to exhibit characteristics of intelligence needs to have deductive as well as inductive abilities. With respect to the latter, there are theories that explain why induction is important for artificial intelligence; for instance, Valiant's probably approximately correct semantics of learning (Valiant 1984, 2008). Nevertheless, in the case of the former we have a problem first observed by Hintikka (1973), which can be stated in the following way:

1. A deduction is valid if, and only if, the conjunction of its premisses, says $\phi_1, \ldots, \phi_n$, implies its conclusion, ψ.
2. In this case, $\phi_1 \wedge \cdots \wedge \phi_n \rightarrow \psi$ is a tautology, i.e., valid deductions are equivalent to propositions without information.
3. Therefore, deductions are uninformative.

A. Beraldo de Araújo (✉)
Center for Natural and Human Sciences (CCNH), Federal University of ABC (UFABC), São Bernardo do Campo, SP, Brazil
e-mail: anderson.araujo@ufabc.edu.br

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, Synthese Library 376, DOI 10.1007/978-3-319-26485-1_9

This was called by Hintikka the *scandal of deduction*. It is a scandal not only because it contradicts Aristotle's maxim that deductions are important for obtaining beliefs, but, mainly, in virtue of the fact that we actually obtain information via deductions. Due to this and other reasons, Floridi has proposed a theory of strong semantic information in which semantic information is true well-formed data (cf. Floridi 2004). From this standpoint, Floridi is capable of explaining why some logical formulas are more informative than others. If we want to explain why deductions, not only propositions, are important for knowledge acquisition and intelligent processing, we cannot, however, apply Floridi's theory. The main reason is that it was designed to measure the static semantic information of the data expressed by propositions. In contrast, knowledge acquisition and intelligence are dynamic phenomena, associated in some way to the flow of information.

In the present work, we propose to overcome that limitation by defining a measure of semantic information in Floridi's sense, but in the context of a dynamic perspective about the logical features of databases associated to valid deductions (Sect. 9.2). We restrict ourselves to first-order deductions and adopt a semantic perspective about them, which means that deductions are analyzed in terms of structures. There are two reasons for that choice. The first one is that the scandal of deduction is usually approached in terms of structures associated to valid deductions (cf. Sequoiah-Grayson 2008). In other words, it is a problem with the semantic informativity of deductions. The second reason is technical: databases are finite structures and so there is, in general, no complete deductive first-order logical system for finite structures (cf. Ebbinghaus and Flum 1999).

We impose to ourselves the methodological constraint that a good approach to semantic informativity must be applicable to real computational systems. More specifically, we look for a solution that enables us to link semantic information and artificial intelligence. Because of that, we propose to measure the degree of semantic informativity of deductions as a dynamic phenomenon, based on certain explicit definitions of insertions and deletions on databases. In that context, the concepts of coherency and relevancy are explained by the operations of insertion and deletion (Sect. 9.3), and so semantic informativity is defined in terms of relevancy and coherency (Sect. 9.4). This approach leads us to a solution to the scandal of deduction (Sect. 9.4). Moreover, using straightforward definitions, our definition of semantic informativity provides an immediate heuristic principle to improve the deductive power of computational systems in semantic terms.

9.2 Databases

We intend to analyze the semantic informativity obtained via deductions. According to Floridi, semantic information is true well-defined data (Floridi 2011). As far as logic is concerned, we can say that data is in some way expressed by propositions. In

general, deductions are compounded of two or more propositions. Thus, we need to consider databases, because they are just organized collections of data (cf. Kroenke and Auer 2007). From a logical point of view, the usual notion of database can, in its turn, be understood in terms of the mathematical concept of structure.

Definition 9.1. A *database* is a pair $D = (A, T)$ where A is a finite first-order structure over a signature S and T is a correct (all propositions in T are true in A) finite first-order theory about A.

Example 9.1. Let $D_1 = (A_1, T)$ be a database with signature $S = (\{s, l, a\}, \{C, E, H\})$, for $s = \ulcorner\text{São Paulo}\urcorner$, $l = \ulcorner\text{London}\urcorner$, $a = \ulcorner\text{Avenida Paulista}\urcorner$, $C = \ulcorner\text{City}\urcorner$, $E = \ulcorner\text{Street}\urcorner$ and $H = \ulcorner\text{To have}\urcorner$, such that $A_1 = (\{\bar{s}, \bar{l}, \bar{a}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$ and $T = \{\forall x(Cx \rightarrow \exists y Hxy), \forall x(Cx \vee Ex), \neg El, Cs\}$.

Remark 9.1. In the Example 9.1 we have used $\ulcorner\alpha\urcorner = \beta$ to mean that the symbol β is a formal representation of the expression α. Besides, X_β is the interpretation of β in the structure A and we use a bar above letters to indicate individuals of the domain of A.

The fact that T is correct with respect to A does not exclude, however, the possibility that our database does not correspond to reality. In the Example 9.1, it is true in A that $Hla \wedge Ea$; in words, it is true in A that London has a street called Avenida Paulista, which, until date of the present paper, it is not true. The theory T represents the fundamental facts of the database that are took as true, that is to say, they are the *beliefs* of the database. It is important to observe that T may not be complete about A, it is possible that not all true propositions about A are in T; Example 9.1 shows this.

We turn now to the dynamics of changes in databases that will permit us to measure semantic informativity. As a general principle, we establish that these changes in the structure of databases must preserve the true propositions of their theories through operations that we call *structural operations*. The first structural operation is the action of putting possibly new objects in the structure of the database and, then, interpreting a possibly new symbol in terms of these objects.

Definition 9.2. Let $D = (A, T)$ be a database over a signature S. An *insertion* of the n-ary symbol $\sigma \in S'$ in D is a database $D' = (A', T)$ where A' is a structure over $S' = S \cup \{\sigma\}$ with the following properties:

1. $A'(\tau) = A(\tau)$ for all $\tau \neq \sigma$ such that $\tau \in S$;
2. If $n = 0$, then $A' = A \cup \{a\}$ and $A'(\sigma) = a$, provided that, for all $\phi \in T$, $A' \vDash \phi$;
3. If $n > 0$, then $A' = A \cup \{a_1, \ldots, a_n\}$ and $A'(\sigma) = A(\sigma) \cup \{(a_1, \ldots, a_n)\}$, provided that, for all $\phi \in T$, $A' \vDash \phi$.

Example 9.2. Let $D_1 = (A, T)$ be the database of the Example 9.1. The database $D_2 = (A_2, T)$ with signature $S' = S \cup \{b\}$, where $b = \ulcorner\text{Shaftesbury Avenue}\urcorner$, and $A_2 = (\{\bar{s}, \bar{l}, \bar{a}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \bar{a}_b, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$ is an insertion of

b in D. On the other hand, $D_3 = (A_3, T)$ is an insertion of E in D_3 where $A_3 = (\{\bar{s}, \bar{l}, \bar{a}, \bar{b}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \bar{a}_b, \{\bar{s}, \bar{l}\}_C, \{\bar{a}, \bar{b}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$ is an S'-structure. Nonetheless, for $A^* = (\{\bar{s}, \bar{l}, \bar{a}, \bar{b}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \bar{b}_b, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$, an S'-structure, we have that $D^* = (A^*, T)$ is not an insertion of b in D_1 because in this case $A^* \not\models \forall x(Cx \vee Ex)$.

The Example 9.2 shows that it is not necessary to introduce a new object in the structure of the database to make an insertion (cf. database D_2); it is sufficient to add a possibly new element in the interpretation of some symbol. On the other hand, it also shows that it is not sufficient to introduce a new object in the structure of the database to make an insertion (cf. database D^*); it is necessary to guarantee that the beliefs of the database are still true in the new structure.

The second structural operation is the action of removing possibly old objects in the structure of the database and, then, interpreting a possibly new symbol in terms of the remaining objects in the database.

Definition 9.3. Let $D = (A, T)$ be a database over a signature S. A *deletion* of the n-ary symbol $\sigma \in S'$, $S - \{\sigma\} \subseteq S' \subseteq S$, from D is a database $D' = (A', T)$ where A' is a structure over S' with the following properties:

1. $A'(\tau) = A(\tau)$ for all $\tau \neq \sigma$ such that $\tau \in S$;
2. If $n = 0$, $A - \{A(\sigma)\} \subseteq A' \subseteq A$ and $A'(\sigma) \in A'$, provided that, for all $\phi \in T$, $A' \vDash \phi$;
3. If $n > 0$, $A - \{a_1, \ldots, a_n\} \subseteq A' \subseteq A$ and $A'(\sigma) = A(\sigma) - \{(a_1, \ldots, a_n)\}$, provided that, for all $\phi \in T$, $A' \vDash \phi$.

Example 9.3. Let $D_1 = (A, T)$ be the database of the Example 9.1. The database $D'_2 = (A'_2, T)$ with signature S and $A'_2 = (\{\bar{s}, \bar{l}, \bar{a}\}, \bar{a}_s, \bar{l}_l, \bar{a}_a, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$ is a deletion of s from D. On the other hand, $D'_3 = (A'_3, T)$ is a deletion of H from D'_2 where $A'_3 = (\{\bar{s}, \bar{l}, \bar{a}\}, \bar{a}_s, \bar{l}_l, \bar{a}_a, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{l}, \bar{a})\}_H)$ is a S-structure. Nonetheless, for $A'_4 = (\{\bar{l}, \bar{a}\}, \bar{a}_s, \bar{l}_l, \bar{a}_a, \{\bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{l}, \bar{a})\}_H)$, an structure over the signature S, we have that $D'_4 = (A_4, T)$ is not a deletion of C from D'_2 because in this case, in despite of $A'_4 \vDash \phi$ for $\phi \in T$, we have that $D'_4(H) \neq D'_2(H)$. Note, however, that D'_4 is a deletion of C from D'_3.

Example 9.3 illustrates that the restriction $A - \{a_1, \ldots, a_n\} \subseteq A' \subseteq A$ means that we can delete at most the elements of the domain that we remove from the interpretation of the symbol under consideration.

Insertions and deletions on databases are well known primitive operations (cf. Kroenke and Auer 2007). Nevertheless, to the best of our knowledge, they have being thought of as undefined notion. Here we have proposed, however, a logical perspective about databases and we have defined explicitly the operations of insertion and deletion to analyze the importance of semantic information. In Araújo (2014), a more strict notion of structural operation is given.

9.3 Coherency and Relevancy

In this section, we propose a dynamic perspective about coherency and relevancy. This approach will permit us to evaluate how many structural operations a proposition requires to become true. We will use these concepts to define the semantic informativity in the next section.

Definition 9.4. An *update* $\bar{D}$ of an S-database D is a finite or infinite sequence $\bar{D} = (D_i : 0 < i \leq \omega)$ where $D_1 = D$ and each D_{i+1} is a insertion or deletion in D_i. An update $\bar{D}$ of D is *coherent* with a proposition ϕ if $\bar{D} = (D_1, D_2, \ldots, D_n)$ and $A_n \vDash \phi$; otherwise, $\bar{D}$ is said to be *incoherent* with ϕ.

Example 9.4. Let D_1 be the database of the Example 9.1 and D_2 be the database of the Example 9.2. The sequence $\bar{D} = (D_1, D_2)$ is an update of D coherent with Eb and Hlb. Let D_1 be the database of the Example 9.1 and D'_2, D'_3 and D'_4 be the databases of the Example 9.3. The sequence $\bar{D}' = (D_1, D'_2, D'_3, D'_4)$ is an update of D coherent with $Es \wedge \neg Hsa$ but not with $s = a$ because the last proposition is false in $A'_4 = (\{\bar{l}, \bar{a}\}, \bar{a}_s, \bar{l}_l, \bar{a}_a, \{\bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{l}, \bar{a})\}_H)$.

In other words, an update for a proposition ϕ is a sequence of changes in a given database that produces a structure in which ϕ is true. In this way, we can measure the amount of coherency of propositions.

Definition 9.5. Let $\bar{D} = (D_1, D_2, \ldots, D_n)$ be an update of the database D. If $\bar{D}$ is coherent with ϕ, we define the *coherency* of ϕ with $\bar{D}$ by

$$H_{\bar{D}}(\phi) = \frac{\min\{m \leq n : A_m \vDash \phi\}}{\sum_{i=1}^{m} i}$$

but if $\bar{D}$ is incoherent with ϕ, then

$$H_{\bar{D}}(\phi) = 0.$$

A proposition ϕ is said to be *coherent* with the database D if $H_{\bar{D}}(\phi) > 0$ for some update $\bar{D}$, otherwise, ϕ is *incoherent* with D.

Remark 9.2. In the definition of coherency the denominator $\sum_{i=1}^{m} i$ is used in order to normalize the definition (the coherency is a non-negative real number smaller than or equal to 1).

Example 9.5. The coherence of Eb and Hlb with the update $\bar{D}$ of the Example 9.4 is the same $2/3$, i.e., $H_{\bar{D}}(Eb) = H_{\bar{D}}(Hlb) \approx 0.66$ and so $H_{\bar{D}}(Eb \wedge Hlb) = H_{\bar{D}}(Eb \vee Hlb) \approx 0.66$. On the other hand, with respect to the coherence of Es, $\neg Hsa$ and $\neg s = a$ and with the update $\bar{D}'$ of the Example 9.4, we have $H_{\bar{D}'}(Es) \approx 0.66$, $H_{\bar{D}'}(\neg Hsa) = 0.4$, $H_{\bar{D}'}(s = a) = 0$ and so $H_{\bar{D}'}(Es \wedge \neg Hsa) = 0.4$ but $H_{\bar{D}'}(Es \wedge s = a) = 0$.

The Example 9.5 exhibits that, given an update, we can have different propositions with different coherency, but we can have different propositions with the same coherency as well. The fact that $H_{\bar{D}}(Eb) = H_{\bar{D}}(Hlb) = H_{\bar{D}}(Eb \wedge Hlb) \approx 0.66$ shows that coherency *is not* a measure of the complexity of propositions. It seems natural to think that $Eb \wedge Hlb$ is in a sense more complex than Eb and Hlb. Here we do not have this phenomena. Moreover, the fact that $H_{\bar{D}}(Eb \wedge Hlb) = H_{\bar{D}}(Eb \vee Hlb) \approx 0.66$ makes clear that, since some propositions have a given coherency, many others will have the same coherency. Another interesting point is that $H_{\bar{D}'}(Es) > H_{\bar{D}'}(\neg Hsa)$ but $H_{\bar{D}'}(\neg Hsa) = H_{\bar{D}}(Eb \wedge Hlb) \approx 0.33$. This reflects the fact that updates are sequences. First, we had made Es coherent with $\bar{D}'$, and later $\neg Hsa$ was made coherent with $\bar{D}$. When $\neg Hsa$ is coherent with $\bar{D}'$ there is nothing more to be done, as far as the conjunction $Es \wedge \neg Hsa$ is concerned.

These remarks show that our approach is different from the one given in (cf. D'Agostino and Floridi 2009). It is not an analysis of some concept of complexity associated to semantic information.[1] From this viewpoint we can obtain an important result in the direction of a solution to the scandal of deduction.

Proposition 9.1. *For every database $D = (A, T)$ and update $\bar{D}$ coherent with ϕ, $H_{\bar{D}}(\phi) = 1$ for every ϕ such that $A \vDash \phi$. In particular, for ϕ a tautology in the language of D, $H_{\bar{D}}(\phi) = 1$, but if ϕ is not in the language of D, $0 < H_D(\phi) < 1$. In contrast, for every contradiction ψ in any language, $H_{\bar{D}}(\psi) = 0$ for every update $\bar{D}$ of D.*

In virtue of our focus in this paper is conceptual, we do not provide proofs here (cf. Araújo (2014) for that). By now, we only observe that if a tautology has symbols different from those in the language of the database, it is necessary to make some changes in order to become it true. In contrast, a proposition is incoherent with a database when there is no way to change it in order to become the proposition true. For this reason, contradictions are never coherent.

We turn now to the concept of relevancy. For that, let us introduce a notation. Consider $(\phi_1, \phi_2, \ldots, \phi_n)$ a valid deduction of formulas over the signature S whose premises are in the set $\Gamma = \{\phi_1, \phi_2, \ldots, \phi_m\}$ and its conclusion is $\phi = \phi_n$. We represent this deduction by $\Gamma\{\phi\}$.

Definition 9.6. Let $\bar{D} = (D_1, \ldots, D_n)$ be an update of the S-database $D = (A, T)$ coherent with ϕ. The *relevant premises* of the deduction $\Gamma\{\phi\}$ with respect to $\bar{D}$ are the premises that are true in D_n but are not logical consequences of T, i.e., the propositions in the set $\bar{D}(\Gamma)$ of all $\psi \in \Gamma$ for which $D_n \vDash \psi$ but $T \nvDash \psi$.

[1] In Araújo (2014), we do an analysis of informational complexity similar to the one presented here about coherency, but these two concepts are different. In further works, we will examine the relation between them.

Example 9.6. Let $\bar{D}'' = (D_1)$. Then, $\bar{D}''(\{Ea\}\{\exists xEx\}) = \{Ea\}$. Now let us consider a more complex example. Let $\bar{D} = (D_1, D_2)$ be the update of Example 9.4. In this case, $\bar{D}(\{\forall x(Cx \rightarrow \neg Ex), Cb\}\{\neg Eb\})$ is not defined because $\neg Eb$ is false in $A_2 = (\{\bar{s}, \bar{l}, \bar{a}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \bar{a}_b, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$. Nonetheless, consider the new update $\bar{D}''' = (D_1, D_2, D_3, D_4, D_5)$ such that D_3 is the insertion in Example 9.2, $A_4 = (\{\bar{s}, \bar{l}, \bar{a}, \bar{b}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \bar{b}_b, \{\bar{s}, \bar{l}\}_C, \{\bar{a}, \bar{b}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$ and $A_5 = (\{\bar{s}, \bar{l}, \bar{a}, \bar{b}\}, \bar{s}_s, \bar{l}_l, \bar{a}_a, \bar{b}_b, \{\bar{s}, \bar{l}\}_C, \{\bar{a}\}_E, \{(\bar{s}, \bar{a}), (\bar{l}, \bar{a})\}_H)$. Then, $\bar{D}'''(\{\forall x(Cx \rightarrow \neg Ex), Cb\}\{\neg Eb\}) = \{\forall x(Cx \rightarrow \neg Ex)\}$.

In the definition of relevant premises, we have adopted a semantic perspective oriented to conclusion of deductions: the relevancy of the premises of a deduction are determined according to an update in which its conclusion is true. Example 9.6 illustrates that point, because it is only possible to evaluate the relevancy of $\{\forall x(Cx \rightarrow \neg Ex), Cb\}\{\neg Eb\}$ in an update like $\bar{D}'$ in which the conclusion $\neg Eb$ is true. Another point to be noted is that we have established a strong requirement about what kind of premises could be relevant: the relevant premises are just the non-logical consequences of our believes.

Definition 9.7. Let D be an S-database. If $\bar{D}$ is an update of D coherent with ϕ, the *relevancy* $R_{\bar{D}}(\Gamma)$ of the deduction $\Gamma\{\phi\}$ in $\bar{D}$ is the cardinality of $\bar{D}(\Gamma)$ divided by the cardinality of Γ, i.e.,

$$R_{\bar{D}}(\Gamma) = \frac{|\bar{D}(\Gamma)|}{|\Gamma|},$$

but, if $\bar{D}$ is incoherent with ϕ, then $R_{\bar{D}}(\Gamma) = 0$.

Example 9.7. We have showed in Example 9.6 that $R_{\bar{D}''}(\{Ea\}\{\exists xEx\}) = 1$ and $R_{\bar{D}'''}(\{\forall x(Cx \rightarrow \neg Ex), Cb\}\{\neg Eb\}) = 0.5$.

In the example above, $R_{\bar{D}'''}(\{\forall x(Cx \rightarrow \neg Ex), Cb\}\{\neg Eb\}) = 0.5$ shows us that we can have valid deductions with non-null relevancy in extended languages. Nonetheless, the fact $R_{\bar{D}''}(\{Ea\}\{\exists xEx\}) = 1$ exhibits that it is not necessary to consider extended languages to find deductions with non-null relevancy. Thus, we have a result that will be central to our solution of the scandal of deduction.

Proposition 9.2. *For every database $D = (A, T)$, update $\bar{D}$ of D and deduction $\Gamma\{\phi\}$, if T is a complete theory of A or $\Gamma = \oslash$, then $R_{\bar{D}}(\Gamma) = 0$. In particular, tautologies and contradictions have null relevancy.*

Therefore, deductions can be relevant only when we do not have a complete theory of the structure of the database. Moreover, as deductions, isolated logical facts (tautologies and contradictions) have no relevance. This means that we have at hand a deductive notion of relevancy.

9.4 Semantic Informativity and Artificial Intelligence

Having at hand the dynamic concepts of coherence and relevance, now it seems reasonable to say that the more coherent the conclusion of a valid deduction is the more informative it is, but the more relevant its premises are the more information they provide. We use this intuition to define the semantic informativity of valid deductions.

Definition 9.8. The *semantic informativity* $I_{\bar{D}}(\Gamma\{\phi\})$ of a valid deduction $\Gamma\{\phi\}$ in the update $\bar{D}$ of the database D is defined by

$$I_{\bar{D}}(\Gamma\{\phi\}) = R_{\bar{D}}(\Gamma)H_{\bar{D}}(\phi).$$

The idea behind the definition of semantic informativity of a valid deduction $\Gamma\{\phi\}$ is that $I_{\bar{D}}(\Gamma\{\phi\})$ is directly proportional to the relevance of its premises Γ and to the coherency of its conclusion ϕ. Given $\Gamma\{\phi\}$ and an update $\bar{D}$ of D, if we have $R_{\bar{D}}(\Gamma) = 0$ or $H_{\bar{D}}(\phi) = 0$, then the semantic informativity of $\Gamma\{\phi\}$ is zero, it does not matter how $\Gamma\{\phi\}$ is. Now, if $H_{\bar{D}}(\phi) = 0$, then, by definition, $R_{\bar{D}}(\Gamma) = 0$. Thus, if the computational system, whose database is D, intends to evaluate $I_{\bar{D}}(\Gamma\{\phi\})$ for some update $\bar{D}$, it should look for a $\bar{D}$ coherent with ϕ, i.e., a $\bar{D}$ for which $H_{\bar{D}}(\phi) > 0$. In other words, our analysis of the semantic informativity is oriented to the conclusion of valid deductions – as we did with respect to relevancy.

Example 9.8. Given the updates $\bar{D}''$ and $\bar{D}'''$ of the Example 9.6. Then, $I_{\bar{D}''}(\{Ea\}\{\exists xEx\}) = 1 \cdot 1 = 1$ and $I_{\bar{D}'''}(\{\forall x(Cx \to \neg Ex), Cb\}\{\neg Eb\}) = 0.5 \cdot 5/15 \approx 0.17$.

In the definition of $I_{\bar{D}}(\Gamma\{\phi\})$ the relevancy of the premises, $R_{\bar{D}}(\Gamma)$, is a factor of the coherency of the conclusion, $H_{\bar{D}}(\phi)$. For that reason, if a computational systems intends to evaluate the semantic informativity of a proposition ϕ, it must measure $H_{\bar{D}}(\phi)$ and, then, multiply it by its relevancy $R_{\bar{D}}(\{\phi\})$. Hence, the semantic informativity of a proposition ϕ can be thought of as a special case of the informativity of the valid deduction $\{\phi\}\{\phi\}$.

Definition 9.9. The *semantic informativity* $I_{\bar{D}}(\phi)$ of a proposition ϕ in the update $\bar{D}$ of the database D is defined by

$$I_{\bar{D}}(\phi) = I_{\bar{D}}(\{\phi\}\{\phi\}).$$

Example 9.9. Considering the update $\bar{D}''$ of Example 9.6, we have that $I_{\bar{D}''}(Ea) = I_{\bar{D}''}(\exists xEx) = 1$ but $I_{\bar{D}''}(Ea \to \exists xEx) = 0$. If we consider the update $\bar{D}'''$ of Example 9.6, we have that $I_{\bar{D}'''}((\forall x(Cx \to \neg Ex) \wedge Cb) \to \neg Eb) = 0$, but $I_{\bar{D}'''}(\forall x(Cx \to \neg Ex)) = 1, I_{\bar{D}'''}(Cb) = 0$ and $I_{\bar{D}'''}(\neg Eb) = 0.4$.

Example 9.9 shows that semantic informativity measures how many structural operations we do in order to obtain the semantic information of a proposition. It

is for that reason that $I_{\bar{D}''}(Cb) = 0$, false well-defined data is not semantically informative; it should be true. In other words, it is a measure of semantic information in Floridi's sense (cf. Floridi 2011). From this, we can solve Hintikka' scandal of deduction.

Proposition 9.3. *For every valid deduction* $\psi_1, \ldots, \psi_n \vDash \phi$ *in the language of D,* $I_{\bar{D}}((\psi_1 \wedge \cdots \wedge \psi_n) \rightarrow \phi) = 0$ *for every update* $\bar{D}$*. Nonetheless, if* $\psi_1, \ldots, \psi_n \vDash \phi$ *is not in the language of D,* $I_{\bar{D}}((\psi_1 \wedge \cdots \wedge \psi_n) \rightarrow \phi) > 0$ *for* $\bar{D}$ *coherent with* $(\psi_1 \wedge \cdots \wedge \psi_n) \rightarrow \phi$*.*

This proposition is a solution to the scandal of deduction in two different senses. First, it shows that we can have an informative valid deduction $\{\psi_1, \ldots, \psi_n\}\{\phi\}$ whose associated conditional $\psi_1, \ldots, \psi_n \rightarrow \phi$ is uninformative, for example, the one given in Example 9.8. Second, it shows that it is not completely true that tautologies are always uninformative. When we interpret new symbols, we have some semantic information, notably, the one sufficient to perceive that we have a true proposition – this is a natural consequence of our approach.

In the studies of pragmatics (a linguistics' area of research), Wilson and Sperber formulated two principles about relevant information in human linguistic practice:

> Relevance may be assessed in terms of cognitive effects and processing effort: (a) other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time; (b) other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time. Wilson and Sperber (2004) [p. 608]

If the semantic informativity of propositions cannot be determined by its coherency or relevancy alone, then the two Wilson and Sperber's principles (a) and (b) are in fact parts of one general principle associated to semantic information. Let us put that in precise terms.

Definition 9.10. The *changes* that a proposition ϕ *requires* are the structural operations, insertions and deletions, that generate an update $\bar{D}$ of a given database $D = (A, T)$ coherent with ϕ. A proposition ϕ is *new* if ϕ is not true in A and is not a consequence of the theory T of the database $D = (A, T)$.

Proposition 9.4. *The less changes a new proposition requires, the more informative it is.*

Proposition 9.4 is a direct consequence of Definition 9.9. Therefore, if we show that our Definition 9.9 is not arbitrary, then Proposition 9.4 is not arbitrary too. But Definition 9.9 is not arbitrary. Let us prove that.

Given an update $\bar{D} = (D_1, \ldots, D_n)$ of $D = (A, T)$ and a deduction $\{\phi\}\{\phi\}$, either $R_{\bar{D}}(\{\phi\}) = 0$ or $R_{\bar{D}}(\{\phi\}) = 1$. If $R_{\bar{D}}(\{\phi\}) = 0$, then either $T \vDash \phi$ or $D_n \nvDash \phi$. If $T \vDash \phi$, then there is an update $\bar{D}'$ of D such that $H_{\bar{D}'}(\phi) = 1$, notably, $\bar{D}' = (D)$. If $D_n \nvDash \phi$, then $H_{\bar{D}'}(\phi) = 0$. Finally, if $R_{\bar{D}}(\{\phi\}) = 1$, then $T \nvDash \phi$ as well as $D_n \vDash \phi$ and so $I_{\bar{D}}(\phi) = H_{\bar{D}}(\phi) > 0$. Therefore, we conclude that the relevancy of a proposition does not determine its coherency. On the other hand, if $H_{\bar{D}}(\phi) = 0$,

then $R_{\bar{D}}(\{\phi\}) = 0$, but if $H_{\bar{D}}(\phi) > 0$, this neither necessarily imply that either $R_{\bar{D}}(\{\phi\}) = 1$ nor $R_{\bar{D}}(\{\phi\}) = 0$, because this depends whether $T \vDash \phi$. Hence, we also conclude that the coherency of a proposition does not determine its relevancy. Combining this two conclusions we obtain a general conclusion.

Proposition 9.5. *The semantic informativity of propositions cannot be determined by its coherency or relevancy alone.*

Therefore, Proposition 9.4 is an informational justification for the Wilson and Sperber's principle (b). This is interesting because Wilson and Sperber's principle (b) is an empirical matter under discussion among linguistics (cf. Wilson and Sperber 2004), but here it becomes an informational principle. Using the same strategy, we can also obtain a formal version of Wilson and Sperber's principle (a).

Definition 9.11. If a valid deduction $\Gamma\{\phi\}$ has non-null relevancy in a given update $\bar{D}$ and its conclusion ϕ is new, then the *results* that it *produces* are its relevant premisses and its conclusion, i.e., $\bar{D}(\Gamma) \cup \{\phi\}$, but if ϕ is not new, then the *results* that it *produces* are just its relevant premisses $\bar{D}(\Gamma)$.

Proposition 9.6. *The more results a valid deduction produces, the more informative it is.*

We can, then, combine these two propositions in an schematic one.

Proposition 9.7 (Principle of semantic informativity). *To increase semantic informativity, an intelligent agent, with respect to its database, must perform little changes and produce big results.*

In recent works (cf. Valiant 2008), Valiant have argued that one of the most important challenges in artificial intelligence is that of understanding how computational systems that acquire and manipulate commonsense knowledge can be created. With respect to that point, he explains that some of the lessons from his PAC theory is this:

> We note that an actual system will attempt to learn many concepts simultaneously. It will succeed for those for which it has enough data, and that are simple enough when expressed in terms of the previously reliably learned concepts that they lie in the learnable class. Valiant (2008) [p. 6]

We can read Valiant's perspective in terms of the principle of semantic informativity. The simple propositions are the more coherent propositions, the ones that require small changes in the database. To have enough data is to have propositions sufficient to deduce other propositions and this means that deductions with more results are preferable. Of course, Valiant's remark relies on PAC, a theory about learnability, not on deductivity. It is necessary to develop further works to make clear the relationship between these two approaches. Since we have designed a concept of semantic informativity implementable in real systems, it seems, however, that the possibility of realizing that is open.

9.5 Conclusion

We have proposed to measure the degree of semantic informativity of deductions by means of dynamic concepts of relevancy and coherency. In a schematic form, we can express our approach in the following way:

$$\text{Semantic informativity} \;=\; \text{Relevancy} \;\times\; \text{Coherency}.$$

In accordance with this conception, we showed how the scandal of deduction can be solved. Our solution is that valid deductions are not always equivalent to propositions without information. It is important to note, however, that this problem is not solve in its totality, because here we have analyzed semantic information only from the point of view of relevancy and coherency. Another crucial concept associated to semantic information is the notion of complexity. In Araújo (2014), we study this subject.

We have also derived a principle of semantic informativity that, when applied to computational intelligent systems, shows that an intelligent agent should make few changes in its database and obtain big results. This seems an obvious observation, but it is not. The expressions "few changes" and "big results" here have a technical sense which opens the possibility of relating semantic information and artificial intelligence in a precise way. Indeed, there is a lot of possible developments to be explored, we would like to indicate three.

The first one is to investigate the connections between semantic informativity and machine learning, specially, with respect to Valiant's semantic theory of learning (PAC). As in Valiant's PAC, we can define probability distributions on the possible updates and delineate goals for them – the principle of semantic informativity could play an important role in this point. Moreover, we can also introduce computational complexity constrains to agent semantic informativity. Thus, it will be possible to analyze how many efficient updates (time and space requirements bounded by a function of the proposition size) are necessary for a given proposition to be coherent with the database. In this way, we will be able, for example, to compare the learnability of the concepts which occur in propositions, in the Valiant's sense (cf. Valiant 1984), with respect to their semantic information.

The second possible line of research is to develop a complete dynamic theory of the semantic informativity by incorporating belief revision in the line of AGM theory (cf. Alchourrón et al. 1985). In the present paper, the believes of the database have been maintained fixed, but a more realistic approach should incorporate revision of believes. For example, if we consider distributed systems, the agents probably will have some different beliefs. In this case, it is necessary to analyze the changes of semantic information, conflicting data and so on.

The last point to be explored, but no less important, is to analyze the relationship between our dynamic perspective about semantic information and other static approaches, mainly, with respect to Floridi's theory of strong semantic information

(Floridi 2004). It is important to observe that we have proposed a kind of Hegelian conception about semantic information, according to which semantic informativity is analyzed in semantic terms, whereas, for example, Floridi's conception is Kantian, in the sense that it analyzes the relationship between propositions and the world in order to understand the transcendental conditions of semantic information (cf. Floridi 2011). Our approach seems to be a hegelian turn in the philosophy of information similar to what Brandom did with respect to the philosophy of language (cf. Brandom 1989).

Acknowledgements I would like to thank Viviane Beraldo de Araújo for her support, to Luciano Floridi for his comments on my talk given at PT-AI2013, and to Pedro Carrasqueira for his comments and to an anonymous referee for his (her) criticism on a previous version of this paper. This work was supported by São Paulo Research Foundation (FAPESP) [2011/07781-2].

References

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*, 510–530.

Araújo, A. (2014, forthcoming). A metrics for semantic informativity.

Aristotle. (1989). *Prior analytics* (R. Smith, Trans.). Indianapolis: Hackett Publishing Company.

Brandom, R. (1989). *Making it explicit*. Cambridge: Harvard University Press.

D'Agostino, M., & Floridi, L. (2009). The enduring scandal of deduction: Is propositional logic really uninformative? *Synthese, 167*(2), 271–315.

Ebbinghaus, H., & Flum, J. (1999). *Finite model theory*. Berlin: Springer.

Floridi, L. (2004). Outline of a theory of strongly semantic information. *Minds and Machines, 14*(2), 197–222.

Floridi, L. (2011). *Philosophy of information*. Oxford: Oxford University Press.

Hintikka, J. (1973). *Logic, language games and information. Kantian themes in the philosophy of logic*. Oxford: Clarendon Press.

Kroenke, D., & Auer, D. (2007). *Database concepts*. New York: Prentice Hall.

Sequoiah-Grayson, S. (2008). The scandal of deduction. *Journal of Philosophical Logic, 37*(1), 67–94.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

Valiant, L. (2008). Knowledge infusion: In R. Hariharan & M. Mukund (Eds.), Pursuit of robustness in artificial intelligence. In *FSTTCS 2008*, Bangalore (pp. 415–422).

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 607–632). Malden: Blackwell.

Chapter 10
Information, Computation, Cognition. Agency-Based Hierarchies of Levels

Gordana Dodig-Crnkovic

Abstract This paper connects information with computation and cognition via concept of agents that appear at variety of levels of organization of physical/chemical/cognitive systems – from elementary particles to atoms, molecules, life-like chemical systems, to cognitive systems starting with living cells, up to organisms and ecologies. In order to obtain this generalized framework, concepts of information, computation and cognition are generalized. In this framework, nature can be seen as informational structure with computational dynamics, where an (info-computational) agent is needed for the potential information of the world to actualize. Starting from the definition of information as the difference in one physical system that makes a difference in another physical system – which combines Bateson and Hewitt's definitions, the argument is advanced for natural computation as a computational model of the dynamics of the physical world, where information processing is constantly going on, on a variety of levels of organization. This setting helps us to elucidate the relationships between computation, information, agency and cognition, within the common conceptual framework, with special relevance for biology and robotics.

Keywords Information • Computation • Cognition • Natural computation • Morphological computing • Morphogenesis • Embodied computation

10.1 Introduction

At present we are lacking adequate understanding of cognition in humans (which is what is commonly thought of as "cognition") while at the same time we are trying to develop cognitive robotics and cognitive computing. The contemporary research into artifactual cognition performed in parallel with studies of cognition in humans

G. Dodig-Crnkovic (✉)
Chalmers University of Technology & University of Gothenburg, Gothenburg, Sweden

Mälardalen University, Västerås, Sweden
e-mail: dodig@chalmers.se

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_10

and animals provide us with two-way learning that will result in both better insights in mechanisms of biological cognition and better solutions for cognitive robotics.

In order to study within one framework cognition in living organisms (including humans) and machines (including cognitive software), this article is generalizing some common ideas, thus using extended concepts of <agent>, <observer>, <information>, <computation>, <evolution>, <cognition>, <learning>, and <knowledge>. The basis is the idea of nature as a network of networks of <agents> that exchange information. This generalized type of <agents> exist on the level of fundamental particles, then on the higher level of atoms as composed of networks of elementary particles, then higher still there are molecules consisting of atoms as <agents>. Up in hierarchy of levels of organization of agents there are cells as networks of molecules, organisms as networks of cells, ecologies as networks of organisms, etc. In short there is a fractal structure with recurrent pattern of agents within agents on variety of levels of organization. Dynamics on each level of organization is a result of information exchanges between agents.

<Information> is relational, based on differences, and thus <agent> −dependent. <Agents> are entities capable of acting, that is capable of causing things to happen (elementary particles, atoms, molecules, cells, organisms, etc.). <Computation> is a process of <information> exchange between <agents>, i.e. <information> dynamics or processes on informational structures (Hewitt 2012).

Epistemology is formulated as theory of information (Chaitin 2007), or more specifically as theory of informational structures) (in <cognitive> agents. Even though informational structural realism of Floridi and Sayre is formulated from the perspective of human agents, it is readily generalizable to any other kind of agents processing information from the outside world that guides their organization and behavior.

In this generalized framework, <agents> exist already on a basic physical level, and they form, via processes of self-organization, increasingly complex structures, including living organisms. Living <agents> (all biological systems) are critically dependent on the capability to acquire energy for their own <agency>. Their <cognition> is a property that sustains and governs their process of being alive. Understanding of living <agency> is closely tied to the understanding of origins of life and its evolution.

In present approach we look at evolution as a process that unfolds through morphological <computation>, by morphogenesis and meta-morphogenesis as introduced in Turing (1952) and further studied in Sloman (2013a) and Dodig-Crnkovic (2012d). At different levels of complexity of living <agents>, different levels of <cognitive> information-processing capacities develop – from bacterial colonies consisting of cells with distributed information processing (Xavier et al. 2011) via plants to organisms with nervous systems such as C. elegans (See OpenWorm project http://www.openworm.org that is building the computational model of this microscopic worm) to mammals and finally humans. Organisms preserve <evolutionary memory> as the information about past in their body structures, that in interaction with the environment enable different behaviors. From biological structures as information-processing <machines> we can hope to learn more not

only about the details of form generation (morphogenesis) in biological systems, but also about possible future computational methods and devices inspired by intrinsic natural computation. A lot can be learned from information processing in the brain.

> The uniformity of the cortical architecture and the ability of functions to move to different areas of cortex following early damage strongly suggests that there is *a single basic learning algorithm for extracting underlying structure from richly-structured, high- dimensional sensory data*. (Hinton 2006) (Italics added)

Based on the uniformity of cortical architecture, *Deep Learning Algorithms* have been recently developed as machine learning algorithms for pattern recognition. They are using artificial neural networks that learn in a succession of levels corresponding to increasing levels of abstraction of concepts, with higher-level concepts being defined by means of lower-level ones (Hinton et al. 2006; Hawkins and Blakeslee 2005).[1]

Similarly, based on the behavior of natural systems, Probably Approximately Correct "PAC" algorithms (Valiant 2013) have been proposed as a way of *learning from nature how to learn*. The scope of PAC algorithms is wider than the scope of Deep Learning Algorithms as they offer in general "the unified study of the *mechanisms of evolution, learning, and intelligence* using the methods of computer science". Valiant argues that "to understand the fundamental character of life, learning algorithms are good place to start."

While both PAC algorithms and Deep Learning are centered on *machine learning*, and from there make important connection between (machine) learning algorithms and evolution, I introduced a different path searching for grounding of learning in the mechanisms of <*cognition*> starting with simplest living organisms like bacteria whose processing of information is form of natural computation. Within the framework of info-computationalism, I proposed the unified view of computing nature with <agent>−based fundamental notions of <*information*> and <*computation*> (in the form of information exchanges between <agents>). This builds on Hewitt (2012) who especially focused on interaction and mechanisms of computation as discussed in Dodig-Crnkovic and Giovagnoli (2013). In this approach it is essential that both informational structures and computational processes *appear on variety of levels of organization* (levels of abstraction).

This naturalist strategy aims at explaining human cognitive capacities as a result of evolutionary and developmental processes that we want to model and simulate in order to be able to both better understand humans and other living organisms, how they function and what causes their malfunctions, as well as to learn how to build intelligent computational artifacts that will add to our extended cognition.

[1]The deep learning model (Hinton et al. 2006) involves "learning the distribution of a high level representation using a restricted Boltzmann machine to model each higher layer" (Smolensky 1986).

10.2 Nature as Info-Computation for a Cognizing Agent

In this article I will propose a framework with the aim to naturalize cognition, meaning that we will not study the *concept of cognition* but <cognition> as *natural phenomenon* in any kind of living <agent>.[2] The framework of info-computationalism, presented earlier in Dodig-Crnkovic and Müller (2011); Dodig-Crnkovic (2006). Dodig-Crnkovic and Müller (2011), Dodig-Crnkovic (2006) is based on concepts of <information> and <computation > as two fundamental and mutually dependent concepts defined in a broader sense than what one typically is used to.

Information is understood according to *informational structural realism*) () (as the *structure*, the fabric of the relationships in the universe (for an agent).

Computation is defined as information processing (Burgin 2010) and presents all processes of *changes of the structures of the universe*[3] (natural computationalism or pancomputationalism) (Chaitin 2007; Fredkin 1992; Lloyd 2006; Wolfram 2002; Zuse 1970).[4]

Combining the frameworks of informational structural realism and natural computationalism results in the model of the universe as a computational network with computation defined as the dynamics of natural information, i.e. *natural computation* (Rozenberg et al. 2012). Computing nature represents all structures and processes existing in the physical universe, which necessarily appear in both continuous and discrete form, corresponding to sub-symbolic and symbolic[5] computation levels.

From now on, given the above non-standard definitions I will omit brackets around <agent>, <information>, <computation>, <cognition> etc. and I hope the reader will keep in mind generalized notions that are used in the rest of the article. I will use them only occasionally to emphasize the use of non-standard definition.

A consequence for epistemology for an agent processing information is that *information and reality are seen as one by an agent* (Vedral 2010; Zeilinger 2005), not only in case of humans, but also for other living organisms as cognizing agents (Ben-Jacob et al. 2011; Maturana and Varela 1992; Shapiro 2011). The Relational Nature of Information and Levels of Organization.

---

[2] Some of the issues discussed here have been discussed by the author in a recent book Computing Nature and in the book Information and Computation. This paper presents a synthesis of the previously developed arguments.

[3] This "processing" can be either *intrinsic* (spontaneously going on) within any physical system or *designed* such as in computing machinery.

[4] For majority of computationalists, computing nature is performing discrete computation. Zuse for example represents his calculating space as cellular automata, but the assumption about the type of computation is not essential for the idea that "the universe <computes> its next state from the previous one" (Chaitin 2007).

[5] Sub-symbolic computations take place in neural networks, as signal processing which leads to concept formation following pattern recognition.

Informational structural realism builds on the realist position that the world *exists* independently from observation of cognizing agents. We identify it with *proto-information or potential information*, which is the potential form of existence equivalent to Kant's das Ding an sich. That *potential information* develops into actual information ("*a difference that makes a difference*" (Bateson 1972)) *for a cognizing agent* through interactions that uncover/register[6] aspects of the world.

Hewitt proposed the following general relational[7] definition that subsumes Bateson's definition:

> Information expresses the fact that *a system is in a certain configuration that is correlated to the configuration of another system.* Any physical system may contain information about another physical system. (Hewitt 2007) (Italics added)

Bateson's definition follows from the above formulation if "another system" is an observer for whom the difference in the first system makes a difference. The relational view of information where information is a difference that makes a difference for an agent, can be related to the Wheelers ideas of participatory universe (Wheeler 1990), as well as observer-dependent formulation of endophysics (Rössler 1998) and second-order cybernetics with its observer-dependent knowledge production. "Combining Bateson and Hewitt insights, on the basic level, *information is the difference in one physical system that makes a difference in another physical system, thus constituting correlation between their configurations*" (Dodig-Crnkovic 2014a).

Among correlated systems, of special interest in our discussion of naturalized cognition are *agents – systems able to act* that are capable of causing things to happen, and among agents we will focus on *living agents, that is cognizing agents,* based on Maturana and Varela's understanding that *life is identical with cognition*[8] (Maturana and Varela 1980). In what follows, it should become evident why it is so that all living agents possess some degree of cognition.

The world as it appears (actualizes) for cognizing agents depends on the types of interactions through which they acquire information. "Potential information in the world is obviously much richer than what we observe, containing invisible worlds of molecules, atoms and sub-atomic phenomena, distant objects and similar. Our knowledge about this potential information which reveals with help of scientific instruments continuously increases with the development of new devices and the new ways of interaction, through both theoretical and material constructs" (Dodig-Crnkovic and Müller 2011).

For an agent, potential information actualizes in present time to transform into potential again. Transformations between potential and actual information

[6]The expression "registered" is borrowed from Brian Cantwell Smith (1998).

[7]More on current understanding of information can be found in the Handbook of the Philosophy of Information (Benthem van and Adriaans 2008).

[8]Even though Maturana and Varela identify process of life with cognition, Maturana refuses the information processing view of cognition. It should be noted that it is based on traditional concept of information.

(information process, computation) parallel transformation between potential and kinetic energy. Kampis' component systems (described later on) model information processing in the cell that undergoes cycles of transformations of potential (original informational structure) and actual (current process) in creating new informational structure that is potentiality for a new process. Notions of potentiality and actuality can be traced back to Aristotle, for whom potentiality presents a possibility, while actuality is the change/activity/motion/process that presents realization of that possibility. This relationship parallels being and becoming. Along Aristotle's transitions from potentiality to actuality, we discuss even the transition from actuality to potentiality, closing the cycle of transformations. That would correspond the cycle from original structure (information) via dynamical process (computation) to a new structure (information).

10.3 The Hierarchy of Levels of Natural Information

This article provides arguments for the new kind of understanding, in the sense of (Wolfram 2002), of lawfulness in the organization of nature and especially living systems, emergent from *generative computational laws* of *self-organization based on the concept of agents*. In order to understand the world, organization of the parts in the wholes and interactions between them are crucial. That is where generative processes come in such as *self-organization* (Kauffman 1993) (that acts in all physical systems), and autopoiesis (Maturana and Varela 1980) (that *acts in living cells*).

Self-organization and autopoiesis is effectively described by agent-based models, such as actor model of computation (Hewitt 2012) that we adopt. Given that processes in the cell run in parallel, the current models of parallel computation (including Process calculi, Petri nets, Boolean networks, Interacting state machines, etc.) need to adjust to modelling of biological systems (Fisher and Henzinger 2007).

Interesting framework for information processing in living systems is proposed by Deacon (2011) who distinguishes between the following three levels of natural information (for an agent), as quoted in Dodig-Crnkovic (2012c):

> *Information 1* (Shannon) (data, pattern, signal) (data communication) [syntax]
> *Information 2* (Shannon + Boltzmann) (intentionality, aboutness, reference, representation, relation to object or referent) [semantics]
> *Information 3* ((Shannon + Boltzmann) + Darwin) (function, interpretation, use, pragmatic consequence) [pragmatics]

Deacon's three levels of information organization parallel his three *formative mechanisms*: [Mass-energetic [Self-organization [Self-preservation (semiotic)]]] with corresponding *levels of emergent dynamics*: [Thermo- [Morpho- [Teleo-dynamics]]] and matching *Aristotle's causes*: [Efficient cause [formal cause [final cause]]], according to Dodig-Crnkovic (2012c). Deacon elaborates further that

> Because there are no material entities that are not also processes, and because processes are defined by their organization, *we must acknowledge the possibility that organization itself is a fundamental determinant of physical causality*. At different levels of scale and compositionality, different organizational possibilities exist. And although there are material properties that are directly inherited from lower-order component properties, it is clear that the production of some forms of process organization is only expressed by dynamical regularities at that level. So the emergence of such level-specific forms of dynamical regularity creates the foundation for level-specific forms of physical influence. (Deacon 2011, p. 177)

In the above passage, Deacon expresses the same view that we argue for: matter presents a structure, while causality determines the dynamics of the structure (that we interpret as computation).

In sum, the basic claim of this article is that nature as a network of networks of agents computes through information processes going on in, hierarchically organized layers. Informational structures *self-organize* through intrinsic processes of natural embodied computation/morphological computation as presented in the Introduction to Dodig-Crnkovic and Giovagnoli (2013).

10.4 The Hierarchy of Levels of Physical Computation

If the whole of nature computes, this computation happens on many different levels of organization of the physical matter (Dodig-Crnkovic 2010, 2012b). In Burgin and Dodig-Crnkovic (2011) three *levels of generality of computations* are introduced, from the most general to the most specific/particular one, namely computation defined as the following process:

1. *Any transformation* of information and/or information representation. This leads to natural computationalism in its most general form.
2. *A discrete transformation* of information and/or information representation. This leads to natural computationalism in the Zuse and Wolfram form with discrete automata as a basis.
3. *Symbol manipulation*. This is Turing model of computation and its equivalents.

There are also *spatial levels or scales* of computations (Burgin and Dodig-Crnkovic 2013):

1. *The macrolevel* that includes computations performed by current computational systems in global computational networks and physical computations of macro-objects.
2. *The* microlevel that includes computations performed by integrated circuits.
3. *The nanolevel* that includes computations performed by fundamental parts that are not bigger than a few nanometers. *The molecular* level includes computations performed by molecules.
4. *The quantum* level includes computations performed by atoms and subatomic particles.

For more details see Burgin and Dodig-Crnkovic (2013), presenting the current state of the art on typologies of computation and computational models. By systematization of existing models and mechanisms, the article outlines a basic structural framework of computation.

10.5 Computation on Submolecular Levels

In Hewitt's model of computation, Actors are defined as "the universal primitives of concurrent distributed digital computation". An Actor, triggered by a message it receives, can make local <decisions>, create new Actors, and send new messages (Dodig-Crnkovic 2014b).

> In the Actor Model (Hewitt et al. 1973; Hewitt 2010), computation is conceived as distributed in space, where computational devices communicate *asynchronously* and the entire computation is *not in any well-defined state*. (An Actor can have information about other Actors that it has received in a message about what it was like when the message was sent.) Turing's Model is a special case of the Actor Model. (Hewitt 2012)

The above Hewitt's "computational devices" are conceived as computational agents – informational structures capable of acting, i.e. causing things to happen.

For Hewitt, Actors become Agents only when they are able to "process expressions for commitments including *Contracts, Announcements, Beliefs, Goals, Intentions, Plans, Policies, Procedures, Requests* and *Queries"* (Hewitt 2007). In other words, Hewitt's Agents are human-like or if we broadly interpret the above capacities, life-like Actors. However, we take all Hewitt's Actors to be <agents> with different competences as we are interested in a common framework encompassing all living and artifactual agents.

Hewitt's Actor model (Hewitt 2012) is relational and especially suitable for modeling informational structures and their dynamics. It is based on models of quantum and relativistic physics unlike other models of computation which are based on mathematical logic, set theory, algebra, and similar.

Using actor model, quantum-physical objects such as elementary particles, interacting through force carriers (mediating interactions) can be modeled as actors exchanging messages. The <agency> in simplest physical systems such as elementary particles is exactly their physical capacity to act, to interact and to undergo changes.

As already discussed in Dodig-Crnkovic (2012a), within the framework of info-computationalism, "nature is informational structure – a succession of levels of organization of information for an agent." This structure is different for different agents. Physical reality, das Ding an sich exists as potential information for an agent and actualizes through interactions. This leads to understanding proposed by von Baeyer in his book Information: The New Language of Science (Baeyer von 2004) where he states that "*information is going to replace matter as the primary stuff of the universe*, *providing a new basic framework for* describing and predicting reality".

The current work in quantum physicists on reformulating physics in terms of information, such as proposed by Goyal (2012) in "Information Physics – Towards a New Conception of Physical Reality" and (Chiribella et al. 2012) in "Quantum Theory, Namely the Pure and Reversible Theory of Information" give further support and motivation to info-computational approach. Statistical Thermodynamics can be based on information (or rather lack of information) instead of entropy, as presented in Ben-Naim (2008). In this context it is relevant to mention that even very simple systems can act as <observers>, as described in Matsuno and Salthe (2011) who present possible approach to naturalizing contextual meaning in the case of chemical affinity taken as material agency.

However, if we want tools to manipulate physical systems, such as molecules in the case of studies of origins of life, our tools must be more than theoretical models – they will be *computations "in materio"* as Stepney (2008) called them, explaining: "We are still learning how to use all those tools, both mathematical models of dynamical systems and executable computational models and currently developing 'computation in materio'". That is the reason why physical/natural computing is so important.

10.6 Molecular Computation, Self-Organization and Morphogenesis

Both intrinsic/natural morphogenesis and designed/synthetic/artificial morphogenesis are instructive in the study of evolution and development as embodied computational processes. At present, according to MacLennan "One of the biggest issues that embodied computation faces is the lack of a commonly accepted model of computation" (MacLennan 2010). As morphogenesis seems to have "the characteristics of a coordinated algorithm" it is of special interest to *understand communication patterns of actors in a network* that represents system with morphogenetic dynamics.

> By investigating embryological morphogenesis – a supremely successful example of what we want to accomplish – we can learn many lessons about how communication, control, and computation can be done well at very small scales. (MacLennan 2010)

In the development of an organism, based on the DNA code together with epigenetic control mechanisms, body of a living being is formed on a short time-scale through morphogenesis that governs its development. On a long-time scale, morphological computing governs *evolution* of species. "The environment provides a physical source of biological body of an organism, as a source of energy and matter for its metabolism as well as information, which governs its behavior." According to Dodig-Crnkovic (2008) nervous system and the brain of an organism have evolved "gradually through interactions (computational processes) of simpler living agents with the environment. This process is a result of information self-structuring".

The environment provides an agent with inputs of information and matter-energy, "*where the difference between information and matter-energy is not in kind, but in type of use that organism makes of it*." as argued in Dodig-Crnkovic (2012d). Since "there is no information without representation", (Allo 2008; Landauer 1991) all information is communicated through some physical carrier (light, sound, molecules, ink on a paper, etc.). Thus, the same physical object can be used by an organism as a source of information and a source of nourishment or matter/energy. For example, many organisms use light just as source of information, while other organisms use it in their metabolism as energy source for photosynthesis. Generally, simpler organisms have simpler information structures and processes, simpler information carriers and simpler interactions with the environment. In that sense,

> (B)iotic information is nothing more than the constraints that allows a living organism to harness energy from its environment to propagate its organization. (Kauffman et al. 2008)

According to Maturana and Varela (1980 p. 78), biological autopoetic "machine" is "organized as a network of processes of production, transformation and destruction of components, which through mutual interactions continuously regenerate the network that produced them." Structural coupling with the environment for autopoetic systems is described as continuous dynamical process and considered as an elementary form of *cognition possessed by all life forms*.

Based on the above, we describe cognition as information processing in living organisms, from cellular to organismic level and up to a social cognition. In this framework information is a <substance>, computation is a process and we argue for inseparability of structure/substance and its dynamics/process. If we search for the source of energy necessary to build the constraints and turn environmental resources into the work needed by organisms to run their metabolism, Ulanowicz's process ecology model offers an explanation: "Basically the answer is simply that an aleatoric[9] event took place in which a constraint emerged that allowed a collection of organic molecules to do the work necessary to propagate their organization" (Ulanowicz 2009).

In the spirit of the work of Pfeifer, Lungarella and Sporns (Lungarella and Sporns 2005; Pfeifer et al. 2007), Bonsignorio (2013) studies evolutionary self-structuring of embodied cognitive networks and proposes a framework for the modeling of networks of self-organizing, embodied cognitive agents that can be used for the design of artificial and 'reverse engineering' of natural networks, based on the maximization of mutual information.

Biological systems are networks of interacting parts exchanging information. They are shaped by physical constraints, which also present information for a system, as argued in Kauffman et al. (2008). A living agent is a special kind of <computational> actor that can reproduce and that is capable of undergoing "at least one thermodynamic work cycle" (Kauffman 2000).

[9]Characterized by chance or indeterminate elements, Merriam-Webster online dictionary.

Kauffman's definition (that we adopt) differs from the common belief that (living) agency requires *beliefs* and *desires*, unless we ascribe some primitive form of <belief> and <desire> even to a very simple living agents such as bacteria. The fact is that they act on some kind of <memory> and <anticipation> and according to some <preferences> that might be automatic in a sense that they directly derive from the organisms' morphology (Ben-Jacob 2008, 2009, 2011). Nevertheless bacteria show clear preferences for behaviors that increase organism's survival.

Although the agents capacity to perform work cycles and so persist in the world is central and presents the material basis of life, as Kauffman (2000) and Deacon (2007) have argued, a detailed physical account of it remains to be worked out, and especially relevant in this context is the phenomenon of abiogenesis. Present article is primarily focused on the info-computational foundations of life as cognitive computational taking place on informational structures at different levels of organization of living systems.

10.7 Self-Organization and Autopoiesis

Understanding of cognition as a natural info-computational phenomenon, and reconstruction of the origins, development and evolution of life, can be built on the ideas of self-organisation and autopoiesis.

As described in Dodig-Crnkovic (2014b), the self-organisation as a concept was first defined in the 1960s in general systems theory, and later on in the 1970s in the theory of complex systems. Prigogine was the first to study self-organisation in thermodynamic systems far from equilibrium, which demonstrate *an ability of non-living matter, previously considered to be inert and oppose movement, to self-organize when supplied with energy from the environment*. This process of self-organization is what we describe as a form of morphogenetic/ morphological computing.

Unlike Newtonian laws of motion, which describe *inert matter* that opposes any change of its state of motion, *self-organizing matter* is *active* and spontaneously changes. The ability of inorganic matter to self organize has been studied in Kauffman (1993, 1995). Kauffman's research has inspired investigations into the origins of life that connect the self-organisation of non-living chemical molecules with the abiogenesis and autopoiesis of living beings. Self-organization as fundamental natural process ongoing in all forms of physical systems provides mechanisms for autopoiesis that is characteristic for living organisms.

For our understanding of life as cognition, the work of Maturana and Varela on the basic processes and organization of life is fundamental. They define the process of autopoiesis of a living system as follows:

> An autopoietic machine is a machine organized (defined as a unity) as *a network of processes* of production (transformation and destruction) of components which:

(i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them; and
(ii) constitute it (the machine) as *a concrete unity in space* in which they (the components) exist by specifying the topological domain of its realization as such a network. (Maturana and Varela 1980) p. 78 (emphasis added)

As argued in (Dodig-Crnkovic 2014b) p. 7, "biological systems change their structures and thereby the information processing patterns in a self-reflective, recursive manner through autopoietic processes with structural coupling (interactions with the environment) (Maturana and Varela 1980, 1992). Yet, self-organisation with natural selection of organisms, as a basis for information that living systems build up in their genotypes and phenotypes, is a costly method of <learning> by adapting bodily structures. Higher organisms (which are "more expensive" to evolve) have developed a capability of learning via nervous systems that enable flexible memory with capacity of reasoning as a more efficient way to accumulate knowledge. The transition from "genetic learning" (typical for more primitive forms of life) to the cognitive skills on higher levels of organisation of the nervous system balances the high "production cost" for increasingly complex organisms."

Maturana and Varela claim that the process of autopoiesis that produces and sustains life is the most basic cognitive process.

> Living systems are cognitive systems, and *living as a process is a process of cognition*. This statement is valid for all organisms, with or without a nervous system." (Maturana and Varela 1980) p. 13, emphasis added.

In the info-computational formulation, the process of "life as cognition") () (is understood as information processing in the hierarchy of organizational levels,, starting with molecular networks, via cells and their organizations, to organisms and their networks (Dodig-Crnkovic 2008).

For an agent, the fundamental level of reality is made of proto-information (potential information) that represents the physical structure, while cognition is a process that goes on in real time by information self-structuring (morphological computing) caused by interactions. On a long-time scale it manifests itself as meta-morphogenesis or morphogenesis of morphogenesis. It drives evolution in biological systems, as argued in Sloman (2013b) and Dodig-Crnkovic and Hofkirchner (2011).

In sum, the info-computational model of living organisms connects two basic ideas: self-organization and autopoiesis, generating the sub-cellular, cellular, multicellular, organismic and societal levels of organization. Life processes are supported and constituted by different sorts of morphological computation which on evolutionary time scales define the organisation/structures of living beings, including even processes of meta-morphogenesis (Sloman 2013a).

10.8 Morphological Computing in Component Systems: "Computing in Materio"

Living organisms are described by Kampis as self-modifying systems that must be modeled as "self-referential, self-organizing, "component-systems" (Kampis 1991) which are based on self-generation and self-sustaining (autopoietic) processes and whose behavior, is computational in a general sense, that presents generalization of the Turing machine model. According to Kampis,

> a component system is a computer which, when executing its operations (software) builds a new hardware.... [W]e have a computer that re-wires itself in a hardware-software interplay: the hardware defines the software and the software defines new hardware. Then the circle starts again. (Kampis 1991) p. 223

Living systems are modular and organized in a hierarchy of levels that can be seen as a result of propagation and self-organization of information (Kauffman et al. 2008). A detailed account of the present state of the art of hierarchy of levels/layers can be found in Salthe (2012a, b). Within info-computational framework, levels are informational structure with domain-specific computational modes (intrinsic computation).

An example of a simple biological (component) system, studied in terms of information and computation is described in Xavier et al. (2011) in the following way:

> Thus, each bacterium must be able to sense and communicate with other units in the colony to perform its task in a coordinated manner. The cooperative activities carried out by members of the colony generate a social intelligence, which allows the colony to learn from their environment. In other words, bacterial intelligence depends on the interpretation of chemical messages and distinction between internal and external information. Then a colony can be viewed as a system that analyzes contextual information from its environment, generates new information and retrieves information from the past.

This agrees with the results of Ben-Jacob (2008, 2009, 2011). Talking about grand challenges in the research of natural computing, (Maldonado and Gómez Cruz 2014; Nunes de Castro et al. 2011) identify the central aim of this field to model and harness the above information processes as natural computation.

10.9 Cognition as Cellular Morphological Computation

As a consequence of the received view that only humans possess cognition and knowledge, the study of the cognition in other organisms is still in its beginnings and the origins of <cognition> in first living agents is not well researched. However, if we adopt Maturana and Varela's generalized view of cognition, there are different levels of <cognition> and we have strong reasons to ascribe simpler types of <cognition> to other living organisms. Even such apparently simple organism as bacteria *"sense the environment and perform internal information processing*

(according to the internally stored information) to extract latent information embedded in the complexity of their environment. The latent information is then converted into usable or "active" information that affects the bacterium activity as well as intracellular changes." (Ben-Jacob 2009) Surprisingly perhaps as Pombo et al. (2012) shows, also plants possess memory (encoded as a bodily structure) and ability to learn (adapt, change their morphology) so that they can be said to possess rudimentary forms of cognition.

As already mentioned, autopoiesis (Maturana and Varela 1980) is considered to be the most fundamental level of cognition that is present even in the simplest living organisms. Through evolution, increasingly complex organisms have developed that are able to survive and adapt to their environment. Dodig-Crnkovic (2008) argues that organisms are "able to register inputs (data) from the environment, to structure those into information, and to structure information into knowledge. The evolutionary advantage of using structured, component-based approaches such as data – information – knowledge is based on improved response-time and efficiency of cognitive processes of an organism".

All cognition is embodied – which means that all cognition is based on the bodily experience obtained through interaction with the environment. It holds for all cognitive systems including microorganisms, humans and cognitive robots. More complex cognitive agents build their knowledge both upon direct reaction to input information, and on information processes directed by agents own choices based on information stored in agent's memory.

Information and computation (as processing of information) are basic structural and dynamic elements that can be used to describe self-structuring of input data (data → information → knowledge → data cycles). This interactive natural computational process is going on in a cognitive agent in the adaptive interplay with the environment. Information potentially available in the environment hugely exceeds capacities of cognitive process of any agent. Thus living agents have developed strategies to obtain relevant information. For fungi, content of a book presents no information, but they may use a book as a source of energy (food), that is a basis of information-dynamics for their own bodily structures. Similarly, sunlight triggers just energy production by photosynthesis in plants, while in a human it can trigger the reflection about the nuclear fusion in the sun.

There is a continuous development of morphology from simplest living organism's automaton-like structures to most complex life forms elaborate interplay between body, nervous system with brain and the environment (Pfeifer and Bongard 2006). Cognition is based on restructuring of an agent in the interaction with the environment, while restructuring is morphological computing. From bacteria, that organize in colonies and swarms via more complex multi-cellular organisms and finally humans, cells are the basic cognitive units of a hierarchical distributed process of cognition.

10.10 Summary and Conclusions

To conclude, let me sum up the main points. Firstly, it is important to emphasize that info-computational approach relies on *naturalist methods and scientific results* and even though it assigns <cognitive> capabilities to all living beings, those capacities correspond to empirically established behaviors of biological systems such as e.g. bacteria (Ben-Jacob 2008) and thus, it has no connection to *panpsychism,* that is the view that mind fills everything that exists. Info-computationalism is strictly naturalistic understanding based on physics, chemistry and biology and does not make any assumptions about things like a "mind of an electron". Exactly the opposite, it aims at explaining cognition, and subsequently even mind, through entirely natural processes going on in physical/chemical/biological systems of sufficient complexity.

In general, the ideal of this project is *naturalization* of information, computation, cognition, agency, intelligence, etc.

Shortly, within the info-computational framework we start with the following basic elements:

<Information> is "a structure consisting of differences in one system that cause the difference in another system" (Dodig-Crnkovic and Giovagnoli 2013). In other words, <information> is <observer> −relative. This definition is a synthesis of Bateson and Hewitt definitions.

<Computation> is <information> processing i.e. the dynamics of <information> .

Both <information> and its dynamics <computation> exist on various levels of organization or abstraction/resolution/granularity of matter/energy in space/time.

Of all <agents>, i.e. entities capable of acting that are capable of causing things to happen, only living <agents> are characterized by the ability to actively make choices and act on their own behalf to increase the probability of their own continuing existence. This ability to persist as highly complex organization and to act autonomously is based on the use of energy from the environment, as argued by Kauffman and Deacon.

<Cognition> presents living <agency> consisting of all processes that assure living agent's organizational integrity and continuing activity, and it is equivalent with life (Maturana and Varela 1980).

The dynamics of information leads to new informational structures through self-organization of information that is *morphological computation*. Consequently, corresponding to distinct *layers of structural organization* found in nature (elementary particles, atoms, molecules, cells, organisms, societies, etc.) there are distinct *computational processes of self-organization of information* that implement/realize physical laws (MacLennan 2011). This self-organization is the result of the interactions between different agents/actors as nodes in interaction networks on many levels of organization. In this model each type of actors (in themselves informational structures) exchange messages of the form specific for their level of organization (Dodig-Crnkovic and Giovagnoli 2013).

Finally, as Denning (2007) noticed, "*computing is a natural science*" nowadays, and it assimilates knowledge from and facilitates development of natural sciences – from physics and chemistry to biology, cognitive science and neuroscience. The info-computational approach (Dodig-Crnkovic 2014a) can help to reconceptualize cognition as a self-organising bio-chemical process in living agents, emerging from inorganic matter and evolving spontaneously through information self-structuring (morphological computation). This framework can improve understanding and modelling of living systems, which hitherto have been impossible to effectively frame theoretically because of their complexity, within the common naturalist framework (Dodig-Crnkovic and Müller 2011).

References

Allo, P. (2008). Formalising the "no information without data-representation" principle. In A. Briggle, K. Waelbers, & P. A. E. Brey (Eds.), *Proceedings of the 2008 conference on current issues in computing and philosophy* (pp. 79–90). Amsterdam: Ios Press.

Bateson, G. (1972). In P. Adriaans & J. Benthem van (Eds.), *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology* (pp. 448–466). Amsterdam: University Of Chicago Press.

Ben-Jacob, E. (2008). Social behavior of bacteria: From physics to complex organization. *The European Physical Journal B, 65*(3), 315–322.

Ben-Jacob, E. (2009). Bacterial complexity: More is different on all levels. In S. Nakanishi, R. Kageyama, & D. Watanabe (Eds.), *Systems biology – The challenge of complexity* (pp. 25–35). Tokyo/Berlin/Heidelberg/New York: Springer.

Ben-Jacob, E., Shapira, Y., & Tauber, A. I. (2011). Smart bacteria. In L. Margulis, C. A. Asikainen, & W. E. Krumbein (Eds.), *Chimera and consciousness. Evolution of the sensory self*. Cambridge/Boston: MIT Press.

Ben-Naim, A. (2008). *A farewell to entropy: Statistical thermodynamics based on information*. Singapore/London/Hong Kong: World Scientific.

Bonsignorio, F. (2013). Quantifying the evolutionary self-structuring of embodied cognitive networks. *Artificial Life, 19*(2), 267–289.

Burgin, M. (2010). *Theory of information: Fundamentality, diversity and unification* (pp. 1–400). Singapore: World Scientific Pub Co.

Burgin, M., & Dodig-Crnkovic, G. (2011). Information and computation – Omnipresent and pervasive. In *Information and computation* (pp. vii–xxxii). New York/London/Singapore: World Scientific Pub Co Inc.

Burgin, M., & Dodig-Crnkovic, G. (2013). Typologies of computation and computational models. *Arxiv.org*, *arXiv:1312*.

Cantwell Smith, B. (1998). *On the origin of objects*. Cambridge, MA: MIT Press.

Chaitin, G. (2007). Epistemology as information theory: From Leibniz to Ω. In G. Dodig Crnkovic (Ed.), *Computation, information, cognition – The nexus and the liminal* (pp. 2–17). Newcastle: Cambridge Scholars Pub.

Chiribella, G., D'Ariano, G. M., & Perinotti, P. (2012). Quantum theory, namely the pure and reversible theory of information. *Entropy, 14*, 1877–1893.

Deacon, T. (2011). *Incomplete nature. How mind emerged from matter*. New York/London: W. W. Norton & Company.

Denning, P. (2007). Computing is a natural science. *Communications of the ACM, 50*(7), 13–18.

Dodig-Crnkovic, G. (2006). *Investigations into information semantics and ethics of computing* (pp. 1–33). Västerås: Mälardalen University Press.

Dodig-Crnkovic, G. (2008). Knowledge generation as natural computation. *Journal of Systemics, Cybernetics and Informatics, 6*(2), 12–16.
Dodig-Crnkovic, G. (2010). In J. Vallverdú (Ed.), *Biological information and natural computation*. Hershey: Information Science Reference.
Dodig-Crnkovic, G. (2012a). Info-computationalism and morphological computing of informational structure. In P. L. Simeonov, L. S. Smith, & A. C. Ehresmann (Eds.), *Integral biomathics. Tracing the road to reality*. Berlin/Heidelberg: Springer.
Dodig-Crnkovic, G. (2012b). Information and energy/matter. *Information, 3*(4), 751–755.
Dodig-Crnkovic, G. (2012c). Physical computation as dynamics of form that glues everything together. *Information, 3*(2), 204–218.
Dodig-Crnkovic, G. (2012d). The info-computational nature of morphological computing. In V. C. Müller (Ed.), *Theory and philosophy of artificial intelligence* (SAPERE, pp. 59–68). Berlin: Springer.
Dodig-Crnkovic, G. (2014a). Info-computational constructivism and cognition. *Constructivist Foundations, 9*(2), 223–231.
Dodig-Crnkovic, G. (2014b). Modeling life as cognitive info-computation. In A. Beckmann, E. Csuhaj-Varjú, & K. Meer (Eds.), *Computability in Europe 2014* (LNCS, pp. 153–162). Berlin/Heidelberg: Springer.
Dodig-Crnkovic, G., & Giovagnoli, R. (2013). *Computing nature*. Berlin/Heidelberg: Springer.
Dodig-Crnkovic, G., & Hofkirchner, W. (2011). Floridi's open problems in philosophy of information, ten years after. *Information, 2*(2), 327–359.
Dodig-Crnkovic, G., & Müller, V. (2011). A dialogue concerning two world systems: Info-computational vs. mechanistic. In G. Dodig Crnkovic & M. Burgin (Eds.), *Information and computation* (pp. 149–184). Singapore/Hackensack: World Scientific.
Fisher, J., & Henzinger, T. A. (2007). Executable cell biology. *Nature Biotechnology, 25*(11), 1239–1249.
Fredkin, E. (1992). Finite nature. *Proceedings of the XXVIIth Rencotre de Moriond*, Les Arcs, Savoie, France.
Goyal, P. (2012). Information physics – Towards a new conception of physical reality. *Information, 3*, 567–594.
Hawkins, J., & Blakeslee, S. (2005). *On intelligence*. New York: Times Books, Henry Holt and Co.
Hewitt, C. (2007). What is commitment? Physical, organizational, and social. In P. Noriega, J. Vazquez-Salceda, G. Boella, O. Boissier, & V. Dign (Eds.), *Coordination, organizations, institutions, and norms in agent systems II* (pp. 293–307). Berlin/Heidelberg: Springer.
Hewitt, C. (2010). Actor model for discretionary, adaptive concurrency. *CoRR*, *abs/1008.1*. Retrieved from http://arxiv.org/abs/1008.1459
Hewitt, C. (2012). What is computation? Actor model versus Turing's model. In H. Zeni (Ed.), *A computable universe, understanding computation & exploring nature as computation*. Singapore: World Scientific Publishing Company/Imperial College Press.
Hewitt, C., Bishop, P., & Steiger, P. (1973). A universal modular ACTOR formalism for artificial intelligence. In N. J. Nilsson (Ed.), *IJCAI – Proceedings of the 3rd International Joint Conference on Artificial Intelligence* (pp. 235–245). Standford: William Kaufmann.
Hinton, G. (2006). To recognize shapes, first learn to generate images, *UTML TR 2006–004*.
Hinton, G., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation, 18*, 1527–1554.
Kampis, G. (1991). *Self-modifying systems in biology and cognitive science: A new framework for dynamics, information, and complexity* (pp. 1–564). Amsterdam: Pergamon Press.
Kauffman, S. (1993). *Origins of order: Self-organization and selection in evolution*. New York: Oxford University Press.
Kauffman, S. (1995). *At home in the universe: The search for laws of self-organization and complexity*. New York: Oxford University Press.
Kauffman, S. (2000). *Investigations*. New York/London: Oxford University Press.

Kauffman, S., Logan, R., Este, R., Goebel, R., Hobill, D., & Shmulevich, I. (2008). Propagating organization: An enquiry. *Biology and Philosophy, 23*(1), 27–45.
Landauer, R. (1991). Information is physical. *Physics Today, 44*, 23–29.
Lloyd, S. (2006). *Programming the universe: A quantum computer scientist takes on the cosmos*. New York: Knopf.
Lungarella, M., & Sporns, O. (2005). Information self-structuring: Key principle for learning and development. In *Proceedings of 2005 4th IEEE Int. Conference on Development and Learning* (pp. 25–30).
MacLennan, B. J. (2010). Morphogenesis as a model for nano communication. *Nano Communication Networks, 1*(3), 199–208.
MacLennan, B. J. (2011). Artificial morphogenesis as an example of embodied computation. *International Journal of Unconventional Computing, 7*(1–2), 3–23.
Maldonado, C. E., & Gómez Cruz, A. N. (2014). Biological hypercomputation: A new research problem in complexity theory. *Complexity*, *wileyonline* (1099–0526). doi:10.1002/cplx.21535.
Matsuno, K., & Salthe, S. (2011). Chemical affinity as material agency for naturalizing contextual meaning. *Information, 3*(1), 21–35.
Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht/Holland: D. Reidel Pub. Co.
Maturana, H., & Varela, F. (1992). *The tree of knowledge*. Boston: Shambala.
Nunes de Castro, L., Silveira Xavier, R., Pasti, R., Dourado Maia, R., Szabo, A., & Ferrari, D. G. (2011). The grand challenges in natural computing research: The quest for a new science. *International Journal of Natural Computing Research (IJNCR), 2*(4), 17–30.
Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think – A new view of intelligence*. Cambridge, MA: MIT Press.
Pfeifer, R., Lungarella, M., & Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science, 318*, 1088–1093.
Pombo, O., Torres, J. M., & Symons J, R. S. (Eds.). (2012). *Special sciences and the unity of science (Logic, Epi.)*. Berlin/Heidelberg: Springer.
Rössler, O. (1998). *Endophysics: The world as an interface*. Singapore/London/Hong Kong: World Scientific.
Rozenberg, G., Bäck, T., & Kok, J. N. (Eds.). (2012). *Handbook of natural computing*. Berlin/Heidelberg: Springer.
Salthe, S. (2012a). Hierarchical structures. *Axiomathes, 22*(3), 355–383.
Salthe, S. (2012b). Information and the regulation of a lower hierarchical level by a higher one. *Information, 3*, 595–600.
Shapiro, J. A. (2011). *Evolution: A view from the 21st century*. New Jersey: FT Press Science.
Sloman, A. (2013a). Meta-morphogenesis. Retrieved from http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-morphogenesis.html
Sloman, A. (2013b). Meta-morphogenesis: Evolution and development of information-processing machinery. In S. B. Cooper & J. van Leeuwen (Eds.), *Alan Turing: His work and impact* (p. 849). Amsterdam: Elsevier.
Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 194–281). Cambridge, MA: MIT Press.
Stepney, S. (2008). The neglected pillar of material computation. *Physica D: Nonlinear Phenomena, 237*(9), 1157–1164.
Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London, 237*(641), 37–72.
Ulanowicz, R. E. (2009). *A third window: Natural life beyond Newton and Darwin*. West Conshohocken: Templeton Foundation Press.
Valiant, L. (2013). *Probably approximately correct: Nature's algorithms for learning and prospering in a complex world*. New York: Basic Books.
van Benthem, J., & Adriaans, P. (2008). *Philosophy of information*. Amsterdam: North Holland.

Vedral, V. (2010). *Decoding reality: The universe as quantum information* (pp. 1–240). Oxford: Oxford University Press.

von Baeyer, H. (2004). *Information: The new language of science*. Cambridge, MA: Harvard University Press.

Wheeler, J. A. (1990). Information, physics, quantum: The search for links. In W. Zurek (Ed.), *Complexity, entropy, and the physics of information*. Redwood City: Addison-Wesley.

Wolfram, S. (2002). *A new kind of science*. Wolfram Media. Retrieved from http://www.wolframscience.com/

Xavier, R. S., Omar, N., & de Castro, L. N. (2011). Bacterial colony: Information processing and computational behavior. In *Nature and biologically inspired computing (NaBIC), 2011 Third World Congress on*, pp. 439–443, 19–21 Oct 2011. doi: 10.1109/NaBIC.2011.6089627. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6089627&isnumber=6089255

Zeilinger, A. (2005). The message of the quantum. *Nature, 438*(7069), 743.

Zuse, K. (1970). *Calculating space. Translation of "Rechnender Raum"*. Cambridge, MA: MIT Technical Translation.

Chapter 11
From Simple Machines to Eureka in Four Not-So-Easy Steps: Towards Creative Visuospatial Intelligence

Ana-Maria Olteţeanu

Abstract This chapter builds an account of the cognitive abilities and mechanisms required to produce creative problem-solving and insight. Such mechanisms are identified in an essentialized set of human abilities: making visuospatial inferences, creatively solving problems involving object affordances, using experience with previously solved problems to find solutions for new problems, generating new concepts out of old ones. Each such cognitive ability is selected to suggests a principle necessary for the harder feat of engineering insight. The features such abilities presuppose in a cognitive system are addressed. A core set of mechanisms able to support such features is proposed. A unified system framework in line with cognitive research is suggested, in which the knowledge-encoding supports the variety of such processes efficiently.

Keywords Creativity • Problem-solving • Insight • Visuospatial intelligence

11.1 Introduction

We are still far from building machines that match human-like visuospatial intelligence, creative problem solving, or the more elusive trait of insight. Creativity and creative problem-solving have fascinated humans ever since individuals able to wield such skills with great prowess have existed, generating many legends and anecdotes. Thus Archimedes is said to have had the insight of how to measure the volume of a crown while immersing himself in a bathtub (Vitruvius Pollio 1914). Watson has recounted to have dreamt of spiral staircases before settling on the double helix solution for the problem of the structure of DNA. In a speech given at the German Chemical Society, Kekulé mentioned to have day-dreamt an Ouroboros-like snake biting its tail or a tibetan knot before discovering the structure of benzene (Fig. 11.1).

A.-M. Olteţeanu (✉)
Universität Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany
e-mail: amoodu@informatik.uni-bremen.de

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_11

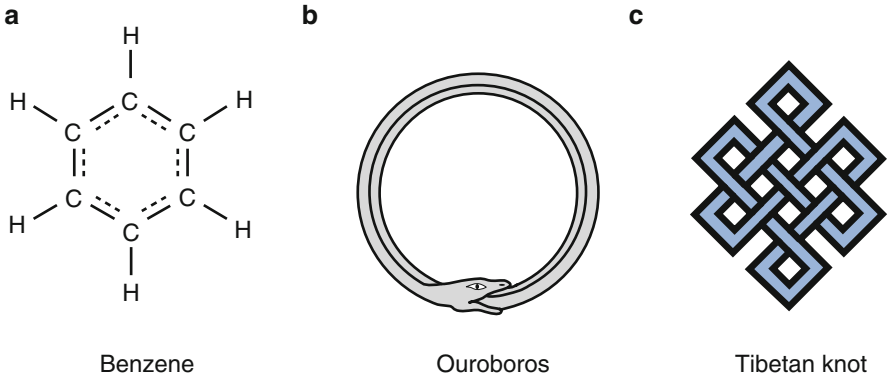


Fig. 11.1 A depiction of Kekulé's day-dream: (**a**) The benzene molecule; (**b**) Ouroboros symbol of a serpent eating its own tail; (**c**) Tibetan knot

These introspective and sometimes second-hand accounts cannot be taken as facts, but as descriptions of various phenomenological experiences of insight in problem-solving (or assumptions about these experiences if the account is second-hand). To discriminate the myth from the fact studies in empirical settings on insight problem-solving and creativity have been employed (Maier 1931; Duncker 1945) and creativity tests developed (Kim (2006) offers a review of one of the most used such tests – TTCT – the Torrance Test of Creative Thinking).

Yet some of these accounts coalesce in their narrative, pointing at similar phenomenology. This invites the question whether such similar phenomenology is the result and indicator of a certain set of cognitive processes or the phenomenological narrative converges because a narrative schema[1] about insight has become ingrained in our culture. If the answer is the former, these accounts might hold some reverse-engineering potential for cognitive and AI scientists, which could work their way back from phenomenological effects,[2] using as a lead the cognitive processes that have generated them, to further decypher the hidden mechanisms of insight.

Phenomenological complications aside, insight and creative problem-solving generally figure amongst the pinnacle of human cognitive abilities (other animals are capable of creative tool use (Köhler 1976) and some analogy-making (Gillan et al. 1981), however we are unaware of any experimental set-up able to test

[1]Several different proposals which aim to summarize all macro-narratives exist, a compelling one being offered by Booker (2004), however for a computational treatment of micro narrative schemas see Chambers and Jurafsky (2010). In the context of insight, the established narrative schema could be about inspiration that comes to the discoverer after a lot of work in a spontaneous flash, in which various parts of the problem are "perceived" together with similar inspiration-conducive objects.

[2]This can hold true only if the imagery which accompanies insight is real and in direct relation to the causal processes of insight – i.e. visual imagery is perceived because visual components of concepts are activated and worked upon with visual and other processes in order to propose a solution.

for insight in animals). Individuals able to produce great leaps of thought seem to have always existed among us (Watson 2005, 2011), yet creative problem-solving is something many normal human beings do on a day-to-day basis – when putting a new mechanism together out of known parts, improvising a tool when lacking one, coming up with new ideas, concepts and strategies, adapting older problem-solving strategies to new situations. Compared to the achievements of other primates or artificial intelligence agents, even the smallest human creative intelligence accomplishments are remarkable.

We define *productive cognition* (cf. Wertheimer 1945) as the general ability to create new knowledge, concepts, tools and objects, mechanisms, theories and systems of thought. The emphasis here is on producing a new object or a solution that has not existed or was not known or experienced before. Research into creativity, creative problem-solving and insight, all aspects of productive cognition, has valuable potential impact for both AI and cognitive science. The engineering applications are related to smarter, more robust AI systems, which can solve tasks in new environments with higher flexibility and an ability to adapt their previous knowledge to the new problems they encounter. Ideally these agents should be able to produce new information (concepts, theories, new relevant relations, hypotheses on how to represent problems), the usability of which can then be tested by classical computational paradigms. The benefits for cognitive science are in what the computational modeling of and experimentation with such abilities can tell us about how they function in their natural state in human cognition.

For artificial systems, creative problem-solving poses a high complexity challenge, bringing about the question how new types of knowledge and hypotheses can be created that are actually useful, other than by logical inference. For cognitive science, the issue is rather what kind of representations and processes enable the functioning of such abilities. These two questions connect and this chapter deals with them in tandem.

A unified framework (Newell 1994) aimed at exploring and implementing cognitively-inspired creative problem-solving and insight is proposed. Here the scientific interest is focused on determining what kind of knowledge representation-processing pairs can generally support a variety of creative problem-solving processes with more ease than previous computational paradigms. To determine such types of knowledge representation and processes, an essential set of cognitive abilities and the features they presuppose in a cognitive system is analysed. These abilities each illuminate a different cognitive mechanism needing implementation in order to reach higher abilities in productive systems. The way all these mechanisms can be integrated to participate in the higher-level abilities of creative problem-solving and insight is then shown. Furthermore, the framework is constructed on an initial visuospatial inference ability, which if replicated should account at the problem-solving level for similar phenomenological effects as Watson's and Kekulé's accounts.

The rest of this chapter is organised as follows. Section 11.2 gives a flavor of the matters which have preoccupied researchers in the various aspects of productive cognition: creativity, creative problem-solving and insight, thus presenting the

issues involved in the construction of a creative problem-solving framework. Section 11.3 defines such a framework in four steps, elaborating on the cognitive features which need implementation and proposing knowledge organization and knowledge processing mechanisms, in a path from visuospatial intelligence to insight. Section 11.4 concludes the chapter with a discussion about the cognitive abilities the system presupposes as essential, a birds-eye view about how the mechanisms that are proposed at each level interact, and future work required to implement, test and refine this theoretical framework.

11.2 Aspects of Productive Cognition: Creativity, Creative Problem-Solving and Insight

11.2.1 Creativity and Creative Problem-Solving

Boden (2003) distinguishes between historical creativity (h-creativity), which produces results original on the scale of human history, and psychological creativity (p-creativity), which yields contributions that are creative from individual perspective. She further differentiates between combinatorial and exploratory-transformational creativity. Combinatorial creativity is a form of producing new, unusual combinations or associations out of known ideas. Exploratory-transformational creativity is an exploration of variations, and changes to/restructuring of the conceptual space. As the term *conceptual space* is not very clearly defined (Ritchie 2001; Wiggins 2001), its compatibility with uses by others (Gärdenfors 2004) is hard to determine.

Another lens through which creative processes are approached is that of the difference between convergent and divergent thought (Guilford 1967). Convergent thought is assumed to employ previously known reasoning strategies, familiar heuristics and data, as to arrive to an accurate, logical solution. By contrast, divergent thought is a search for many different potential solutions, with various degrees of correctness, where the emphasis is on production of a diversity of possible solutions, not on accuracy. Thus divergent thought is assumed to be creative and associative in nature, exploring multiple possible solutions and courses of action, and evaluating them in a quick and rough manner. Such solutions don't need to be logical or traditionally used heuristics – they can be associationist in nature, using previous knowledge from different fields to enable what is popularly described by the term of "leaps of thought". However, this categorization is rather abstract, with each category being able to contain many processes, and creative problem-solving is the type of endeavor which assumes both abilities – a divergent stage to find possible different solutions, and a convergent one to follow through the consequences of such solutions.

Implicit processes are generally considered to play an important role in creative problem-solving, with some models focusing on explicit-implicit process

interaction (Hélie and Sun 2010). The incubation stage in insight is considered to be a process which takes place under conscious awareness. However, the relationship between the concepts of divergent thought, implicit processing and the incubation stage has not been clearly disseminated in the literature (though one can assume some degree of overlap).

Important roles in creativity are played by analogy (Holyoak and Thagard 1996) and metaphor (Lakoff and Johnson 1980, 1999). Both analogy and metaphor are generally considered to be processes of transferring knowledge from a known field (source) to a less known field (target), with various purposes, like: enriching the unknown field, having some starting assumptions and knowledge to test, explaining that field to a learner in a fashion which is connected with knowledge that the learner already possesses as to allow for a quick comprehension start in the new field, aesthetical effects (with comprehension consequences).

An important aspect that any theory of creativity needs to account for is concept generation or composition. Concept formation literature in its various forms (prototype theory – Rosch (1975), exemplar theory – Medin and Shoben (1988), theory theory – Murphy and Medin (1985)) has not traditionally dealt with aspects of concept composition. More recently theories have been proposed on this matter (Aerts and Gabora 2005; Fauconnier and Turner 1998). The latter, a conceptual blending account, proposes that various elements and relations from different scenarios or concepts are blended in an unconscious process, as to produce new concepts. This account finds its ancestry in Arthur Koestlers concept of bisociation of matrices (Koestler 1964).

Concept discovery (Dunbar 1993) and restructuring (possibly linked to Boden's transformational creativity processes of restructuring the conceptual space) are an important feature in other creative cognition activities – scientific discovery (Nersessian 2008; Langley 2000; Klahr and Dunbar 1988) and technological innovation (Thagard 2012).

The essential difference between creativity and creative problem-solving seems to be one of evaluation type. Creativity is not enough to problem-solve, as an emphasis is put on the utility of the solution, or of the new knowledge and exploration forms (conceptual tools, ideas) created in the problem-solving process. Ultimately, the aesthetic and originality value of a creative solution fades in front of its utility or lack thereof.

This adds hardship in the construction of such a system, but helps in the evaluation process. The constraints bring about the benefit that utility is measured with more ease than aesthetic value and even originality. However, a system's ability to propose new solutions, hypotheses or approaches towards a problem, which might not ultimately work in practice but are valid proposals with chances of success from a human perspective, is a good enough criterion for satisficing creative problem-solving demands.

A specific though challenging kind of creative problem-solving which might shed some light on the cognitive mechanisms at work is insightful problem-solving.

11.2.2 The General Problem of Insight

In the context of Boden's taxonomy (Boden 2003), two types of insight can be determined – a p-creative one (finding the representation which can lead to solving a problem that has been previously solved by others) and a h-creative one (finding a new solution or problem-representation altogether, a case found in the realm of scientific discovery and technological innovation). In order to address the issue of knowledge organization and processes a machine would need to possess to be able the have insight the way humans do, we will focus here on red thread features generally associated with insight (of both kinds). We will however address this distinction again in Sect. 11.3.4.

Encyclopaedia Britannica defines (insight 2014) as:

> immediate and clear learning or understanding that takes place without overt trial-and-error testing. Insight occurs in human learning when people recognize relationships (or make novel associations between objects or actions) that can help them solve new problems

In Sternberg and Davidson (1996) insight is:

> suddenly seeing the problem in a new way, connecting the problem to another relevant problem/solution pair, releasing past experiences that are blocking the solution, or seeing the problem in a larger, coherent context

One example of an insight problem which has been studied in empirical settings is the candle problem (Duncker 1945). The participant is given a box of thumbtacks, a book of matches and a candle. The task is to fix the lit candle on a wall so that the candle wax won't drip onto the table below. The participants give various solutions, including attaching the candle with a thumbtack to the wall, or glueing it with part of the wax. The traditional correct solution to this problem is to use the box of thumbtacks as a platform for the candle, and attach it to the wall using one of the thumbtacks. The accuracy and speed of the participants in solving this problem increases when the box of thumbtacks is presented empty, with the thumbtacks out. A possible reason for this is that participants find it harder to see the box as a platform while its affordance as a container is already used (through the box being full).

In Maier's classical two string problem (Maier 1931), the participants are put in a room which has two strings hanging from the ceiling. Their task is to tie the two strings together. It is impossible to reach one string while holding the other. However, various objects are scattered across the room. The traditional solution to this problem is to use a heavy object (normally the pliers), attach it to one of the strings, then set that string in a pendular motion. Finding this solution can be triggered by the experimenter touching the string, thus making salient its motion affordance and directing the subjects to think of the string as a pendulum.

The literature on insight generally uses a four-stage process proposed by Wallas (1926). The four stages are: familiarization with the problem, incubation (not thinking about the problem consciously), illumination (the moment of insight) and verification (checking if the solution actually works in practice). Whether the

illumination phase presupposes sudden or incremental problem-solving processes is still debated, and various researchers insist on the importance of various stages. A good general set of characteristics for insight problems is proposed by Batchelder and Alexander (2012):

1. They (insight problems) are posed in such a way as to admit several possible problem representations, each with an associated solution search space.
2. Likely initial representations are inadequate in that they fail to allow the possibility of discovering a problem solution.
3. In order to overcome such a failure, it is necessary to find an alternative productive representation of the problem.
4. Finding a productive problem representation may be facilitated by a period of non-solving activity called incubation, and also it may be potentiated by well-chosen hints.
5. Once obtained, a productive problem representation leads quite directly and quickly to a solution.
6. The solution involves the use of knowledge that is well known to the solver.
7. Once the solution is obtained, it is accompanied by a so-called "aha!" experience.
8. When a solution is revealed to a non-solver, it is grasped quickly, often with a feeling of surprise at its simplicity, akin to an aha! experience.

The main challenge in replicating insight in artificial systems is that insight problems are not search problems in the traditional (Newell and Simon 1972) sense. The problems are ill-structured (Newell 1969) for a classical search-space type of solving, and defining an appropriate representation is part of the solution (cf. Simon 1974). For humans, this is the point where functional fixedness gets in the way – with solvers getting stuck in representation types which are familiar and sometimes seem implied by the problem, but are actually inappropriate. Thus a machine replicating such phenomena will have to be able to do some form of metareasoning and re-representation.

11.3 A Framework for Creative Problem-Solving Based on Visuospatial Intelligence

Various work relates visuospatial intelligence to the creation of abstract concepts, and to the process of abstract thought in general (Mandler 2010; Freksa 1991, 2013). Thus Mandler proposes that complex abstract concepts are built developmentally on top of already acquired spatial concepts. This would explain the pervasiveness of spatial templates (Lakoff and Johnson 1980, 1999) in human metaphor, spatial priming influences (Tower-Richardi et al. 2012) and shape bias (Landau et al. 1988; Imai et al. 1994). In his analysis of 100 scientific discoveries (Haven 2006) and 100 technological innovations (Philbin 2005), Thagard (2012) draws the conclusion that 41 out of the 100 scientific discoveries involve visual representation (spatial

representations is unaccounted for in this analysis though some references are made to kinesthetic ones), with the figure rising to 87 out of the 100 in the technological innovations category. A cognitive architecture which proposes the use of spatio-analogical representations (Sloman 1971) in the modeling of human spatial knowledge processing, without linking them to creative problem-solving is Schultheis and Barkowsky (2011).

The framework proposed here takes into account the importance of visuospatial representations and processes, starting from the general hypothesis that analogical representations and visuospatial (and structure-oriented) processes can be a good representation-process pair for the recognition, manipulation and modification of structures and relation-sets which is necessary in creative problem-solving. Such mechanisms might also offer a bridge over the explanatory gap towards introspective imagery phenomenology which sometimes accompanies moments of insight or creative problem-solving. The rest of this section sets to explore whether abstract creative problem-solving mechanisms can indeed build on simple visuospatial inference mechanisms.

11.3.1 Step 1: Visuospatial Inference

In his paper, Sloman (1971) talks about "formalisation of the rules which make non-linguistic, non-logical reasoning possible". He also gives an example (Fig. 11.2) of visuospatial inference. Figure 11.2a shows a mechanism made of two levers and a pulley, the upward arrow being an indication of the initial direction of motion. The reader should have no problem in visually inferring how the motion propagates through the system, as to arrive at the result represented in Fig. 11.2b.

For a human with cultural familiarity with pulleys and levers, such an inference is visually very simple. However, we take its simplicity for granted, as it comes from our complex visual system's ability to anticipate the motion of objects which it has already learned. Visuospatial inference seems simple because it is a native feature of the human visual system.

To be able to replicate such an ability in artificial intelligence terms, one would have to implement some of the properties of the cognitive visual system that humans generally take for granted. This problem could be translated in AI terms, by giving an artificial system a subset of the six simple machines of antiquity (levers, wheel and axle, pulleys, inclined planes, wedges and screws), together with visuospatial and motor knowledge about each of them. The system could be asked to perform a qualitative assessment of what a machine assembled out of some random set of these components will do – the way the motion will propagate via the so assembled mechanism. This is but a step away from Sloman's example as it involves visuospatial inference with multiple parts and can be thought of as a perceptual task.

This task can be solved by a system via a form of perceptual simulation. One can encode motion affordances together with the pattern recognizer for each specific object in such a way that seeing a certain simple machine shape triggers the

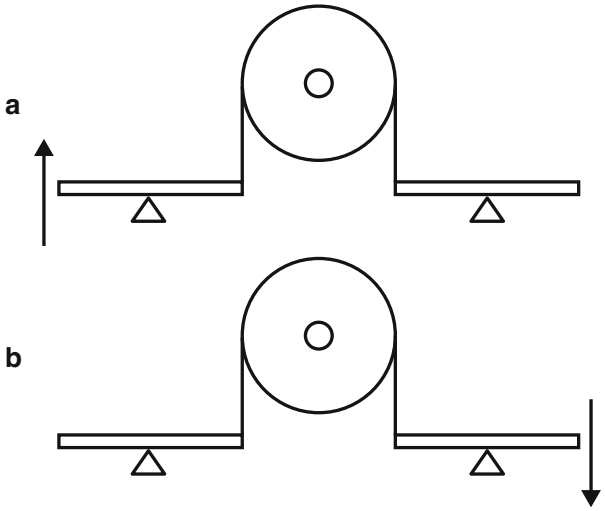


Fig. 11.2 Sloman's diagram of two levers and a pulley: (**a**) motion onset and (**b**) inference result

anticipation or simulation of motion in the artificial system. Whether the simulation of the entire motion is necessary, or just the beginning and end result of such motion can be accessed (once encoded) is something to be settled by cognitive empirical investigation. The system then needs to be endowed with qualitative rules on how motion propagation works between objects which are in contact, and the various ways in which motion changes, or (allowed to learn from) motor simulations of such transitions.

However, as mentioned before, this can be thought of as perceptual inference. In order to talk about problem-solving, two other tasks can be given to the system, in the same problem context:

– to put together a machine starting from a set of known components as to propagate motion in a desired way (multiple solution possible)
– given a set of fixed components, to add missing components so that the mechanism performs a certain type of motion at the end. The number of missing components can be specified or unspecified, however they will be produced out of the system's memory of known machines.

Such problem-solving can rely on the same perceptual simulation (complete or partial) and rules of motion transfer (thus can be entirely visuospatial). In fact it could be a learning trial-and-error process of compositionally adding objects together and checking their motion affordances. The compositionality features allow for objects to be thought of both in terms of simple machines and new composed machines with varying motion affordances.

The implementation of such a system will solve motion anticipation problems with simple machines, and compositionality problems with simple or composed

machines based on their capacity for motion. Besides having interesting features for visuospatial reasoning (maybe an equivalent for a " block-world" classical problem setting), this problem sets the scene for the next steps towards creative problem-solving and insight in a variety of interesting ways. It deals with simple compositionality and decompositionality of objects: an object can be made of various atomic simple machine parts and different compositionality structure can mean different motion affordance, therefore the structure of assembly is essential. The problem can allow for multiple solutions from the part of the solver, and it requires use and manipulation of previous knowledge structures. It is solved based on affordance and compositionality. These features are primitives which we will relate to in the next steps.

11.3.2 Step 2: Creative Use of Affordance

Humans are used to perceiving the world not just in terms of motion anticipation, but also in terms of the affordances (Gibson 1977) that various objects can offer a user, depending on the task at hand. Knowledge of affordances can be considered as a part of commonsense knowledge which displays cultural aspects (as various cultures can be more accustomed to certain objects or tools than others). The cultural variation element does not play a role here, as knowledge of affordances can be treated as a knowledge database which can take whatever form, and thus belong to whatever culture.

Of interest here is the human ability to make creative use of object affordances. When trying to find something to pour liquid in, to carry liquid with and drink from, in conditions in which no cup is available, humans can use a pot, a bucket, or – depending on how desperate the circumstances are – a boot. When wanting to put nails in the wall in the absence of a hammer, humans can use shoes, stones of appropriate size, or other objects. Thus humans can creatively solve problems of the following form: Find an object with a certain affordance, when the object(s) you normally use is not available. Humans can find such objects even if the objects are not normally associated with such an affordance, by speculating on the various properties objects have, and their knowledge about the properties which are normally associated with the affordance.

This type of problem represents a way of making creative inference and use of the affordance properties of objects. In the following we will propose the rough principles of a mechanism for suggesting useful objects in such problems to an artificial agent.

Even simple visuospatial properties such as object shape can lead to inference about affordance. The phenomenon of shape bias (Landau et al. 1988; Imai et al. 1994; Samuelson and Smith 1999), in which children extend names from known to unknown objects based on shape, shows that the human brain considers shape features very important in the context of objects and tools – possibly because of a connection between shape and affordance in these domains. In what follows, we

will propose a mechanism which makes good use of shape in proposing hypotheses, though this can and should be refined to contain more detailed properties which are in a direct relationship with objects' affordances.

To solve such a problem in the spirit of grounded knowledge (Barsalou and Wiemer-Hastings 2005; Barsalou 2003; Gärdenfors 2004), we propose to represent the various objects and tools that the agent knows as distributed concepts. The concepts are distributed over a set of spaces – an affordance space, a visual feature space, and a semantic tag space. Each of these spaces will be organized by similarity, though the similarity metric would be different, depending on what the space contains. Thus visuospatial feature spaces will be organised in terms of feature similarity (of shape, color), verbal tags in terms of semantic or context similarity, affordance spaces in terms of motor trajectories or proprioceptive routines, etc. These spaces could be encoded as self-organised maps (Kohonen 1982). The recognition of an object, or activation of a concept in such a system, would mean the associated activation of points or regions in these spaces (Fig. 11.3). Thus, each concept would be an activation of features over different dimensions, part of which will be more sensory oriented (e.g. the visual features spaces), more functionally and bodily oriented (affordance and motor spaces), and more knowledge oriented (the semantic spaces). Such a cognitive concept could be triggered in a variety of ways: (i) via the semantic tag (its name), the activation of which would spread energy in

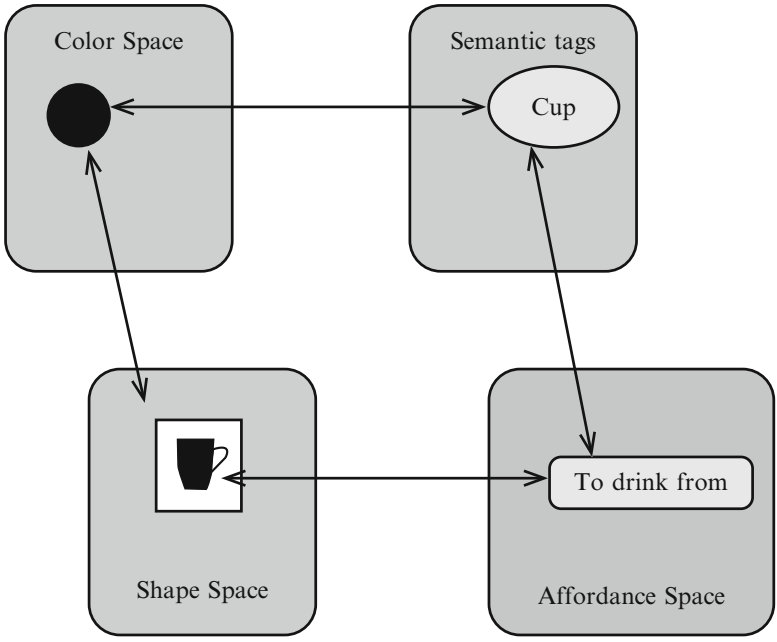


Fig. 11.3 Activation of the concept "cup" over two visuospatial feature spaces, a semantic tag space and an affordance space

the other direct links (how the object looks like, what functions does it normally perform), (ii) via vision input, or (iii) a query related to the affordance which is required.

We prefer such a type of knowledge encoding because the meaning of a concept in such a system becomes grounded in the feature maps, affordance spaces and semantic spaces which we are using (the symbolic paradigm which assumes the meaning is in the verbal form of the concept is refused). However, the proposed mechanism is a hybrid mechanism, as a concept can be interpreted as a symbol (where a symbol is a collection of features, grounded in subsymbolic processes).

Such knowledge organization is useful in two ways. One is that the concept can be activated in different ways, with the entire knowledge network retained about it becoming active. Navigation between such different types of knowledge about one object is possible in natural cognitive systems. Such activation also implies a second benefit, that comes from the encoding in similarity-based maps – navigation between different encoded concepts based on different types of similarity. Thus, when a cup is the direct activation for a certain type of affordance, other neighborhood object shapes are activated as well, with the new object being able to act as a creative substitute, though it might not constitute a traditional solution, nor the type of object the user normally applies in such circumstances.

When a request is made to the system to find an object that is required for a certain affordance (Fig. 11.4), the system will first activate the corresponding

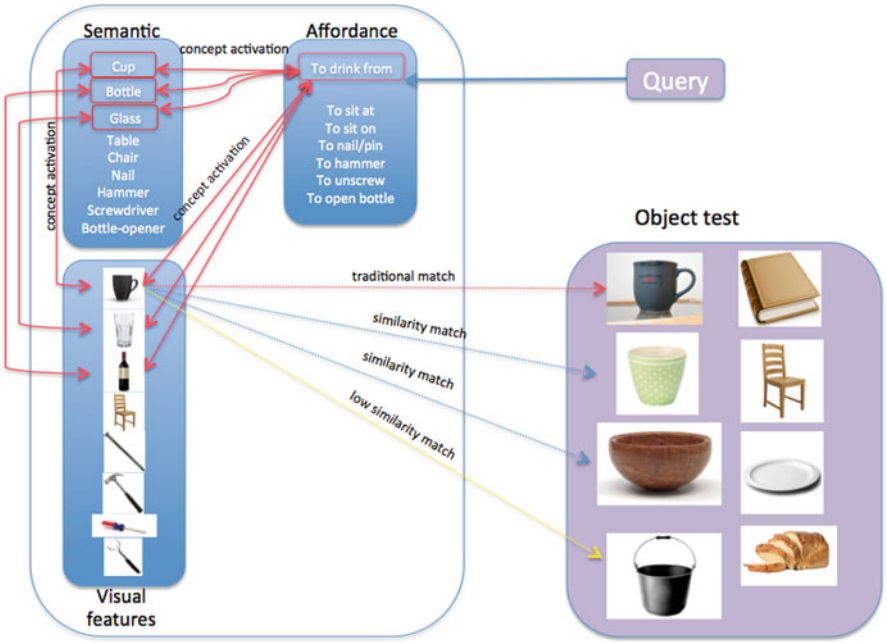


Fig. 11.4 Affordance-based system in action

concepts linked to that affordance (the request is in fact the activation), with the most familiar objects receiving the highest activation. Then the solution object(s) can be searched for in the environment visually. When such a search fails, the threshold of the search drops, and the object will search for something of a similar shape to the familiar solution-object (by quantitative (Forsyth and Ponce 2003) or qualitative (Falomir et al. 2013) means) or/and to objects which are encoded closely to that object in its shape knowledge map. Thus, creative solutions which are not what one set out to search for exactly but can fulfill the function nonetheless can be obtained with limited knowledge, in a visual manner.

11.3.3 Step 3: Concept Generation and Structure Transfer

The third step in this quest for visuospatial creative problem-solving and insight is treated here in two parts. Part (a) deals with the generation of new concepts, and part (b) with problem structure transfer.

11.3.3.1 Step 3A: Generation of New Concepts

Humans can make analogies, use metaphors (Lakoff and Johnson 1980, 1999), blend concepts (Fauconnier and Turner 1998) and sometimes put together features of previously known concepts to create entirely new concepts (like *meme*, *impressionism* or *recursive*), or invent entirely new objects from previously accessible parts or elements.

In making analogies, an important role seems to be played by concept structure and ability to compare and structurally align source and target domain (Gentner 1983, 2010). Spatial schemas are proposed by Lakoff and Johnson (1999) as a process of metaphor creation.

In the proposed framework, the relevant cognitive properties required for concept generation are:

- Associativity of similar concepts on various feature spaces (previously explained in step 2),
- Ability to map a structure in a different feature space, and
- Ability to build concepts compositionally.

In what follows, a few visuospatial processes which make concept generation possible in an artificial system are proposed.

The first process consists of using a previously observed visuospatial relation as a template. Consider an artificial intelligence system that has encoded the relation "chaining" as a visuospatial object – starting from the analogical representation of a chain. The relations encoded in the analogical representation of the object can be used as a template for other units than chain links. First the relations could be extrapolated to similarly shaped objects – like a hoop of string and a scissor's eye

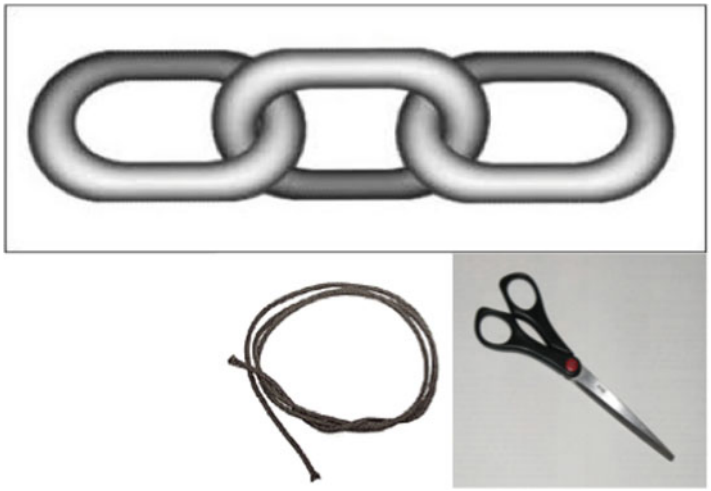


Fig. 11.5 Use of analogical representation of a "chain" as a template for the "chaining" relation

(Fig. 11.5). In the proposed framework, due to visual similarity in the knowledge encoding, such an inference would be natural. The system would thus propose to extend the previous relation, using its template, to other visually similar objects. Such inferences will hold only part of the time, but this is an example of productive reasoning (reasoning which creates a new arrangement of objects in this case), and of transforming the visuospatial analogical representation of a concept into a template for new object arrangements.

Though initially applied to objects with similar visual features, this particular template-relation can be applied at various levels of abstraction, up to concepts like "chaining of events".

Of course not all abstract concepts are derived from visuospatial analogical representation. The point here is to show how some can be derived, as an analogical representation is a very economic way to store relations, and can be used as a structural template for creating new concepts.

A second process that can be used for concept generation is compositionality over such templates. Thus, take the two-tuple relation "bigger-than" as observed in or learned from an example of two trees (Fig. 11.6). An artificial system could match it by size or shape similarity, or use the principle of chaining to a second bigger than relation of two other trees. Via compositionality, this would lead to a three (or more) tuple relation, which would lead to a representation of the concept of "growth". This visuospatial representation of the concept could be used as a template again, via the first process, and adapted to a variety of different domains.

Some such visuospatial datastructures could be compressed to a small subset of features which are consistent across templates, like an upward arrow and a definition of contour (Fig. 11.7). This could be used as an iconic compressed trigger that can activate the concept and stand in for it.

Fig. 11.6 Compositionality of relation – from "bigger-than" to "growth"

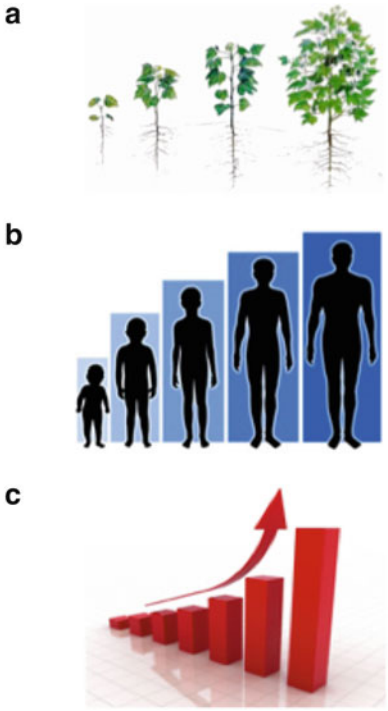


Fig. 11.7 Transition from (**a**) growth template representation, which is (**b**) adapted to a different concept and (**c**) compressed

Thus analogical visuospatial representations can act as mechanisms for concept generation by being used as templates or in applying compositional principles based on similarity principles or visual routines. Such a system could keep track of relations between such templates based on its own experiences with concept generation in a visuospatial semantic map (where semantic is to be understood as meaning relations between such templates), and attempt future composition principles based on such learning (or return to a decomposed form where necessary).

Concept blending can also be implemented by treating concepts as distributed structures over feature spaces, in which the two concepts which participate in the blend each contribute in a varying degree to the structure (and positioning on the feature space) of the new concept. However, for the current purposes, the description of creation and proposal of new structures proved to be a more interesting cognitive feat.

11.3.3.2 Step 3B: Problem Structure Transfer

Problems like the tower of Hanoi are easily solved by people that have understood their heuristics, no matter the shapes of the objects used in the problem presentation. This leads to the obvious conclusion that people are able to detach heuristics from the surface features of the problem, and understand problems in terms of their structure and the heuristics that apply to various structures. General heuristics, like means-end analysis or divide and conquer, are routines which can be deployed independent of the domain. However, the surface features of a problem do play a role in problem-solving, certain problems being solved with much more ease when presented in a certain visual form than in isomorphic but different feature forms (Zhang 1997).

Thus, for humans, a case can be made for both the importance of problem structure, and the importance of surface features in problem-solving efficiency. A system constructed in this proposed framework could deal with both, as follows. A solved problem can be encoded as a distributed structure over the objects and concepts the problem contained (at a lower level their features), the algorithmic steps that have been taken (at a lower level affordances or succesions of motor routines) and the various relationships that have been established during this solving. In the case of a new problem, similar on enough of the encoded properties above, the system could trigger via a form of pattern-completion the previous problem structure – and attempt a subset of similar steps or relation-formation.

The structure could also be elicited in a more direct fashion via remarking upon structural similarities between the problem at hand and a previously solved problem (not on features of the participating objects), or on sets of relations which are common to both. In both cases, the structure of a previously solved problem would thus be transferred to the problem at hand. In case objects of the problem solution or structure are missing, objects and concepts with similar affordances can be used (due to the ability to de-chunk the problem offered by distributed representations).

The essential points in knowledge organization for problem structure transfer are thus threefold:

- It requires the encoding various problem-structures together with their respective component elements and problem-solving procedures (set of affordances, algorithm)
- The ability to match problems to previously known problem-structures and their solutions
- The ability to decompose or recompose problems, as to use different structure-affordance pairs

The last point is further tackled in the issue of insight, when one problem representation structure is not enough.

11.3.4 Step 4: Insight Revisited

As previously discussed, insight is a problem of re-representation, such problems are not solvable via normal search spaces, and their solving doesn't seem to proceed in a step-wise fashion: unlike in non-insight problems, the problem-solvers cannot predict their level of progress or their closeness to the solution (Metcalfe and Wiebe 1987).

In insight problems, it is as if finding the right problem representation is the solution itself. A good representation affords insight directly, by providing the solver with the ability to make the inferences which will lead to the solution. It is thus assumed here that in such problems a form of metareasoning or meta-search happens over the representational structures which can be fitted to the problem, in order to find the one which most obviously affords the (inferences towards the) solution.

In many insight problems, the main problem is thus finding the right problem structure, which is not the normal problem structure that will be ellicited by the objects presented. The various objects participating in the problem have been involved in the commonsense knowledge or can be involved in the commonsense inference of a human being in a variety of problem structures, they posses a variety of affordances. In this framework, insight is defined as a matter of navigating these ellicited structure-affordance pairs until the right one is found (from which further inferences can proceed to a solution). The meta-search space in this framework is richly informed. It encodes the knowledge of the system, together with its similarity metrics over various spaces, and distributed structures with generative compositional properties. The movement in such a search space happens in various dimensions via the similarity of features, context (semantics) and affordance (function) of the distributed objects and templates. This type of knowledge encoding permits informed search via movement through similar structures, or similar objects and the structures they are part of, and creation of new conceptual tools, relations and objects. The right problem structure can be found when searching for a affordance, for similar structures, relations or objects/concept sets.

A different case is that of scientific discovery problems (another variety of the Eureka step), in which it is natural to assume that the "right" problem representation is not in fact found, but created. This framework allows for problem templates to be decomposed, blended, put together, and missing parts to be created out of similarly-affording structures, until a representation is found or created. To close the circle, in the light of the previous steps and the knowledge encoding and processes previously used, solving insight problems (in both forms) becomes somewhat similar to putting simple machines together. The search this time is not one over the known set of simple machines, for the appropriate machine or set of machines to be fitted to the problem of obtaining a certain type of motion or affordance, but for the appropriate problem representation, allowing for compositionality from problem representation fragments, in order to find a problem representation which affords a solution or set of inferences. The motor affordances of the various simple machines are replaced in this case with the affordances the various problem templates can solve.

11.4 Discussion

Productive systems deal in a flexible fashion with the problems they encounter, as to be able to propose new possible solutions based on the knowledge at hand. The framework explored here presupposes a few cognitive properties as being essential for building efficient such productive systems. Efficiency is understood here as computational ease of processing. The proposed framework supports through its knowledge encoding exactly such types of search for a creative solution and re-representation, as to account for cognitive economy principles.

One of these properties is a multidimensional (multisensorial) encoding of concepts (Barsalou 2003), which allows for dynamic memory access based on affordances, visual features or semantic tags. Beside such dynamic access, distributed encoding of concepts allows further grounding in learned, similarity-based organized knowledge and associativity (with traditions in hebbian learning (Hebb 1949), semantic networks (Sowa 1992) and associationism). Such grounding allows for easy navigation of the knowledge space. In a sense, the knowledge space thus becomes the equivalent of a search space in classical problem-solving. However, not all possible states or solutions are mapped. The knowledge encoding merely acts as a map which enables the aforementioned processes to produce more knowledge in a structured, organised manner.

A connected third cognitive property assumed here as essential is flexible, relaxed pattern-recognizing constraints; this allows for non traditional but similar objects to be recognized and accepted as solutions. Essentially this is related to the reality of our imperfect, constructive memory. Though when compared to its machine counterparts a less than optimal part of the human experience, human memory and its imperfections support learning, interpretation and re-interpretation, classification and re-classification, generalization and, by extension

in this framework, creativity and creative problem-solving (rather than a perfect ability to reproduce the things we have perceived with accuracy).

The four steps presented here construct in a coarse manner the necessary abilities of a productive system from the ground up.

- Step one – visuospatial inference – associates visual features, shapes and structures (for the 3D case) with motion affordances, in order to enable motion anticipation in a mechanism composed of simple known parts. This allows simple compositionality principles of affordance, and pattern-fill principles when a small number of objects is given and a mechanism has to be constructed.
- Step two – creative use of affordance – extends the distributed concept encoding, with supplying feature maps organized on similarity principles. This supports a natural search for objects with similar features, affordances or that have been experienced in similar contexts. The flexible threshold in pattern recognition, together with the associativity links enable solutions to be proposed that are not traditional.
- Step (3a) deals with processes for generating new concepts. This creates a conceptual map in which some analogical representations can be used as (1) relation-templates, (2) compositional units that together create new relations, and (3) compression to essentialized visual features. Moreover, the map can keep track of the generative process, and keep relations between the analogical representations which have created new representations through such processes.
- Step (3b) discusses transfer of problem structure into a different problem, based on affordance knowledge (and other possible similarities) of the two structures. This gets closer to the principles of meta-representation, which is attained in step 4. Step (3b) deals with the ability to transfer a set of heuristics, or a problem structure, rather than a small set of relations, that are enclosed in an analogical representation, like in (3a).
- Step 4 puts all the aforementioned principles together. All concepts are grounded in similarity-based maps, where the similarity metric depends on the type of map itself (be it a feature map, an affordance map or a semantic context map). New concepts, conceptual objects and sets of relations are generated as in step (3a), and kept in relations to each other. Problem structures can be transferred in other object spaces. This type of knowledge representation and the aforementioned processes allow for easy re-representation and creation of new problem templates. The framework at this level supports meta-search over known pairs of problem-structure and affordances, enabling the system to find a suitable representation for the problem at hand. If no such problem-structure exists or is known, the knowledge representation and processes allow the system to attempt and create a problem structure compositionally out of known representations.

This framework needs to be implemented and tested. Success criteria of the system are clearly presented in each step: visuospatial inference, creative use of affordance, generation of new concepts, use of problem structure transfer, and solving insight problems, or problems which require creative re-representation. There is a possible difference between the latter two. Insight problems in their reduced form might

require only finding a good representation which affords the solution – though this representation might be quite far away from the natural representation a human would assume for that problem. Problems requiring creative re-representation are closer in kind to scientific discovery, technological innovation, or problems requiring significant change in the conceptual space and tools of the cognitive agent. In these cases, a new representation might need to be created out of known parts, and only once this representation is put together, the parts afford the solution together.

Many of this framework's principles are in line with current cognitive empirical research and theory. However, the cognitive assumptions and and their ensuing implications need to be tested, to see if the framework can hold as a cognitive theory of creative problem-solving, or is a cognitively inspired framework for an artificial intelligence system.

In conclusion, a theoretical framework has been proposed, with a type of knowledge representation and organization meant to support in a unified manner a variety of creative problem-solving abilities and the re-representation features necessary to simulate insightful problem-solving. Each of the various steps has been chosen to underlie an instrumental cognitive ability or mechanism further used in higher level abilities. This theoretical proposal has also been linked to visuospatial types of inference, which might help bridge the gap to the phenomenological experiences of visuospatial insight.

Acknowledgements The work reported in this chapter was conducted in the scope of the project R1-[Image-Space] of the Collaborative Research Center SFB/TR8 Spatial Cognition. Funding by the German Research Foundation (DFG) is gratefully acknowledged.

References

Aerts, D., & Gabora, L. (2005). A theory of concepts and their combinations II: A hilbert space representation. *Kybernetes, 34*(1/2), 192–221.

Barsalou, L. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London, B, Biological Sciences, 358*, 1177–1187.

Barsalou, L., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 129–163). New York: Cambridge University Press.

Batchelder, W. H., & Alexander, G. E. (2012). Insight problem solving: A critical examination of the possibility of formal theory. *The Journal of Problem Solving, 5*(1), 56–100.

Boden, M. (2003). *The creative mind: Myths and mechanisms*. London/New York: Routledge.

Booker, C. (2004). *The seven basic plots: Why we tell stories*. London/New York: Continuum.

Chambers, N., & Jurafsky, D. (2010). A database of narrative schemas. In *Proceedings of the LREC*, Valletta.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17*(3), 397–434.

Duncker, K. (1945). *On problem solving* (Psychological monographs, Vol. 58(5, Whole No.270)). Washington, DC: American Psychological Association.

Falomir, Z., Gonzalez-Abril, L., Museros, L., & Ortega, J. (2013). Measures of similarity between objects from a qualitative shape description. *Spatial Cognition and Computation, 13*, 181–218.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science, 22*(2), 133–187.
Forsyth, D., & Ponce, J. (2003). *Computer vision: A modern approach*. Upper Saddle River: Prentice Hall.
Freksa, C. (1991). Qualitative spatial reasoning. In D. Mark & A. Frank (Eds.), *Cognitive and linguistic aspects of geographic space* (pp. 361–372). Dordrecht/Holland: Kluwer.
Freksa, C. (2013). Spatial computing – how spatial structures replace computational effort. In M. Raubal, D. Mark, & A. Frank (Eds.), *Cognitive and linguistic aspects of geographic space* (Lecture notes in geoinformation and cartography, pp. 23–42). Berlin/Heidelberg/New York: Springer.
Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. Cambridge: Bradford Books.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.
Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*(3), 356–373.
Gibson, J. J. (1977). The theory of affordance. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing*. Hillsdale: Lawrence Erlbaum Associates.
Gillan, D. J., Premack, D., & Woodruff, G. (1981). Reasoning in the chimpanzee: I. analogical reasoning. *Journal of Experimental Psychology: Animal Behavior Processes, 7*(1), 1.
Guilford, J. (1967). *The nature of human intelligence*. New York: McGraw-Hill.
Haven, K. (2006). *100 greatest science inventions of all time*. Westport: Libraries Unlimited.
Hebb, D. (1949). *The organization of behavior*. New York: Wiley.
Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: A unified theory and a connectionist model. *Psychological Review, 117*(3), 994.
Holyoak, K., & Thagard, P. (1996). *Mental leaps: Analogy in creative thought*. Cambridge: MIT.
Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development, 9*(1), 45–75.
insight. (2014). Encyclopaedia britannica online academic edition. http://www.britannica.com/EBchecked/topic/289152/insight.
Kim, K. H. (2006). Can we trust creativity tests? A review of the torrance tests of creative thinking (ttct). *Creativity Research Journal, 18*(1), 3–14.
Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–48.
Koestler, A. (1964). *The act of creation*. New York: Macmillan.
Köhler, W. (1976). *The mentality of apes*. New York: Liveright. (Originally published in 1925).
Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43*, 59–69.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.
Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*(3), 299–321.
Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies, 53*(3), 393–410.
Maier, N. R. (1931). Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology, 12*(2), 181.
Mandler, J. M. (2010). The spatial foundations of the conceptual system. *Language and Cognition, 2*(1), 21–44.
Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology, 20*(2), 158–190.
Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition, 15*(3), 238–246.
Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289.

Nersessian, N. (2008). *Creating scientific concepts*. Cambridge: MIT.
Newell, A. (1969). Heuristic programming: Ill-structured problems. In J. Aronofsky (Ed.), *Progress in operations research, III*. New York: Wiley.
Newell, A. (1994). *Unified theories of cognition*. Cambridge: Harvard University Press.
Newell, A., & Simon, A. (1972). *Human problem solving*. Englewood Cliffs: Prentice Hall.
Philbin, T. (2005). *The 100 greatest inventions of all time: A ranking past and present*. New York: Citadel Press.
Ritchie, G. (2001). Assessing creativity. In *Proceedings of AISB'01 Symposium*, Citeseer.
Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of experimental psychology: General, 104*(3), 192.
Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition, 73*(1), 1–33.
Schultheis, H., & Barkowsky, T. (2011). Casimir: An architecture for mental spatial knowledge processing. *topiCS – Topics in Cognitive Science, 3*, 778–795.
Simon, H. A. (1974). The structure of ill structured problems. *Artificial Intelligence, 4*(3), 181–201.
Sloman, A. (1971). Interactions between philosophy and artificial intelligence: The role of intuition and non-logical reasoning in intelligence. *Artificial Intelligence, 2*(3), 209–225.
Sowa, J. (1992). Semantic networks. In S. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (pp. 1493–1511). New York: Wiley.
Sternberg, R., & Davidson, J. (1996). *The nature of insight*. Cambridge: MIT.
Thagard, P. (2012). Creative combination of representations: Scientific discovery and technological invention. In R. W. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes* (pp 389–405). Oxford/New York: Oxford University Press.
Tower-Richardi, S. M., Brunye, T. T., Gagnon, S. A., Mahoney, C. R., & Taylor, H. A. (2012). Abstract spatial concept priming dynamically influences real-world action. *Front Psychology, 3*, 361.
Vitruvius Pollio, M. (1914). *The ten books on architecture* (pp. 253–254, M. Hicky Morgan, Trans.). Cambridge: Harvard University Press.
Wallas, G. (1926). *The art of thought*. London: Cape.
Watson, P. (2005). *Ideas: A history of thought and invention, from fire to Freud*. New York: HarperCollins.
Watson, P. (2011). *The modern mind: An intellectual history of the 20th century*. London: HarperCollins.
Wertheimer, M. (1945). *Productive thinking*. New York: Harper and Row.
Wiggins, G. A. (2001). Towards a more precise characterisation of creativity in AI. In *Case-Based Reasoning: Papers from the Workshop Programme at ICCBR* (Vol. 1, pp. 113–120).
Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science, 21*(2), 179–217.

Part III
Cognition and Reasoning

Chapter 12
Leibniz's Art of Infallibility, Watson, and the Philosophy, Theory, and Future of AI

Selmer Bringsjord and Naveen Sundar Govindarajulu

Abstract When IBM's Deep Blue beat Kasparov in 1997, Bringsjord (Technol Rev 101(2):23–28, 1998) complained that despite the impressive engineering that made this victory possible, chess is simply too easy a challenge for AI, given the full range of what the rational side of the human mind can muster. However, arguably everything changed in 2011. For in that year, playing not a simple board game, but rather an open-ended game based in natural language, IBM's Watson trounced the best human *Jeopardy!* players on the planet. And what does Watson's prowess tell us about the philosophy, theory, and future of AI? We present and defend snyoptic answers to these questions, ones based upon Leibniz's seminal writings on a universal logic, on a Leibnizian "three-ray" space of computational formal logics that, inspired by those writings, we have invented, and on a "scorecard" approach to assessing real AI systems based in turn on that three-ray space.

Keywords Chess • Watson • Gary Kasparov • IBM • Jeopardy! • Three-ray space • Leibniz

---

The authors are profoundly grateful to IBM for a grant through which the landmark Watson system was provided to RPI, and for an additional gift enabling us to theorize about this system in the broader context of logic, mathematics, AI, and the history of this intertwined trio. Bringsjord is also deeply grateful for the opportunity to speak at PT-AI 2013 in Oxford; for the leadership, vision, and initiative of Vincent Müller; and for the spirited objections and comments received at PT-AI 2013, which served to hone his thinking about the nature and future of (Weak) AI. Thanks are do as well to certain researchers working in the cognitive-computing space for helpful interaction: viz., Dave Ferrucci, Chris Welty, Jim Hendler, Deb McGuiness, and John Licato.

S. Bringsjord (✉)
Rensselaer AI & Reasoning (RAIR) Lab, Department of Cognitive Science, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA

Department of Computer Science, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA
e-mail: selmer@rpi.edu

N.S. Govindarajulu
Rensselaer AI & Reasoning (RAIR) Lab, Department of Cognitive Science, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA
e-mail: naveensundarg@gmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_12

12.1 Introduction

When IBM's Deep Blue beat Kasparov in 1997, Bringsjord (1998) complained that despite the impressive engineering that made this victory possible, chess is simply too easy a challenge for AI, given the full range of what the rational side of the human mind can muster.[1] Systematic human cognition leverages languages and logics that allow progress from set theory to cutting-edge formal physics, and isn't merely focused on an austere micro-language sufficient only to express board positions on an 8×8 grid, a small set of simple rules expressible in a simple set of first-order formulae, and strategies expressible in off-the-shelf techniques and tools in AI.[2] Even techniques for playing invincible chess can be expressed in logical systems well short of those invented to capture aspects of human cognition, and the specific techniques used by Deep Blue were standard textbook ones (e.g., alpha-beta "boosted" minimax search; again, see Russell and Norvig 2009).

However, arguably everything changed in 2011. For in that year, playing not a simple board game, but rather an open-ended game based in natural language, IBM's Watson trounced the best human *Jeopardy!* players on the planet. What is Watson, formally, that is, logico-mathematically, speaking? And what does Watson's prowess tell us about the philosophy, theory, and future of AI? We present and defend snyoptic answers to these questions, ones based upon Leibniz's seminal writings on a universal one, on a Leibnizian "three-ray" space of computational formal logics that, inspired by those writings, we have invented, and on a "score-card" approach to assessing real AI systems based on turn on that three-ray space. It's this scorecard approach that enables a logico-mathematical understanding of Watson's mind. And the scores that Watson earns in turn enables one to predict in broad terms the future of the interaction between *homo sapiens sapiens* and increasingly intelligent computing machines. This future was probably anticipated and called for in no small part by the founder of modern logic: Leibniz. He thought that God, in giving us formal logic for capturing mathematics, sent thereby a hint to humans that they should search for a comprehensive formal logic able to capture and guide cognition across the full span of rational human thought. This—as he put it—"true method" would constitute the "art of infallibility," and would "furnish us with an Ariadne's thread, that is to say, with a certain sensible and palpable medium,

[1] Note that we say: "*rational* side." This is because in the present essay we focus exclusively upon what Leibniz aimed to systematize via his "art of infallibility." Accordingly, in short, we target *systematic* human thought. (In modern terms, this sphere could be defined ostensively, by enumerating what is sometimes referred to as the *formal sciences*: logic and mathematics, formal philosophy, decision theory, game theory, much of modern analysis-based economics (to which Leibniz himself paved the way), much of high-end engineering, and mathematical physics). We aren't concerned herein with such endeavors as poetry, music, drama, etc. Our focus shouldn't be interpreted so as to rule out, or even minimize the value of, the modeling of, using formal logic, human cognition in these realms (indeed, e.g. see Bringsjord and Ferrucci 2000; Bringsjord and Arkoudas 2006); it's simply that in the present essay we adopt Leibniz's focus.

[2] Such as those nicely presented in Russell and Norvig (2009).

which will guide the mind as do the lines drawn in geometry and the formulae for operations which are laid down for the learner in arithmetic."[3] In modern terms, Leibniz would say that God's hint consists in this: The success of purely extensional first-order logic in modeling classical mathematics indicates that more expressive logics can be developed in order to model much more of human cognition.

Our plan for the remainder is this: Next (Sect. 12.2), after recording our agreement with Leibniz that mechanical processing will never replicate human consciousness, we briefly summarize, in four main points, Leibniz's original conception of the art of infallibility, and encapsulate his dream of a comprehensive logico-mathematical system able to render rigorous and infallible all human thinking. In Sect. 12.3, we present our "three-ray" Leibnizian logicist framework for locating a given AI system. Inspired by this three-ray framework, we then (Sect. 12.4) provide a more fine-grained, engineeringish theory of AI systems (the aforementioned scorecard approach), and in Sect. 12.5 use this account to explain what Watson specifically is. We end with some brief remarks about the future of AI.

12.2 Leibniz, Logic, and "The Art of Infallibility"

12.2.1 Disclaimer: Leibniz's Mill and Strong vs. Weak AI

It may be important to inform the reader that at least one of us believes with Leibniz not only that physicalism is false, but believes this proposition for specifically Leibnizian reasons. Leibniz famously argued that because materialism is committed to the view that consciousness consists in mere mechanical processing, materialism can't explain consciousness, and moreover consciousness can't consist merely in the movement of physical stuff. The argument hinges on a thought-experiment in which Leibniz enters what has become, courtesy of this very experiment, a rather famous mill.[4] (The experiment and the anti-physicalist conclusion drawn from it, are defended in Bringsjord 1992.) And he also argued elsewhere, more specifically, that *self*-consciousness is the insurmountable problem for materialism. Note that Leibniz in all this argumentation takes aim at *mechanical cognition*: he argues

[3] From "Letter to Galois," in the year 1677, included in Leibniz and Gerhardt (1890).

[4] In Leibniz's words:

> One is obliged to admit that perception and what depends upon it is inexplicable on mechanical principles, that is, by figures and motions. In imagining that there is a machine whose construction would enable it to think, to sense, and to have perception, one could conceive it enlarged while retaining the same proportions, so that one could enter into it, just like into a windmill. Supposing this, one should, when visiting within it, find only parts pushing one another, and never anything by which to explain a perception. Thus it is in the simple substance, and not in the composite or in the machine, that one must look for perception. (§17 of *Monadology* in Leibniz 1991)

that no *computing machine* can be self-conscious.[5] In other words, Leibniz holds that Strong AI, the view that it's possible to replicate even self-consciousness and phenomenal consciousness in a suitably programmed computer or robot, is false. Note that in our modern terminology, that which is computationally solvable (at least at the level of a standard Turing machine or below) is the same as that which is—as it's sometimes said in our recursion-theory textbooks—*mechanically* solvable.

The previous paragraph implies that the Leibnizian "three-ray" conception of AI that we soon present, as well as our remarks about the future of AI following on that—all of this content we view to be about *Weak* AI (the view that suitably programmed computers/robots can *simulate* any human-level behavior whatsoever), not Strong AI.[6] Put another way, the outward behavior of human persons, when confirmed by a well-defined test, can be matched by a suitably programmed computing machine or robot thereby able to pass the test in question. This specific "test-concretized" form of Weak AI has been dubbed *Psychometric AI* (Bringsjord and Schimanski 2003; Bringsjord 2011; Bringsjord and Licato 2012).

12.2.2 Central Aspects of Leibniz's Dream

While Leibniz is without question the inventor of modern logic, the full details of his original inventions in this regard aren't particularly relevant to our purposes herein. What *is* relevant are four aspects of this invention—aspects that we use as a stepping stone to create and present our aforementioned "three-ray" conception.[7] These four aspects are key parts of Leibniz's dream of a system for infallible reasoning in rational thought. Here, without further ado, is the quartet:

I. Universal Language (**UL**) for Rational Thought Starting with his dissertation, Leibniz sought a "universal language" or "rational language" or "universal characteristic" in which to express all of science, and indeed all of rational thought. (We shall use '**UL**' to refer to this language.) Time and time again he struggled with particular alphabets and grammars in order to render his dream of **UL** rigorous, and while the specifics

[5]Leibniz came strikingly close to grasping *universal* or *programmable* computation. But nothing we say here requires that he directly anticipated Post and Turing. This is true for the simple reason that Leibniz fully understood de novo computation in the modern sense, and also was the first to see that a binary alphabet could be used to encode a good deal of knowledge; and the properties he saw as incompatible with consciousness are those inherent in de novo computation over a binary alphabet $\{0, 1\}$, *and* in modern Turing-level computation.

[6]Bringsjord (1992) is a book-length defense of Weak AI, and a refutation of Strong AI.

[7]With apologies in advance for the pontification, Bringsjord maintains that there is really only one right route for diving into Leibniz on formal logic: Start with the seminal (Lewis 1960), which provides a portal to the turning point in the history of logic that bears a fascinating connection to Leibniz (since Lewis took the first thoroughly systematic move to intensional logic, anticipated by Leibniz). From there, move into and through that which Lewis himself mined, viz. (Leibniz and Gerhardt 1890). At this point, move to contemporary overviews of Leibniz on logic, and into direct sources, now much more accessible in translated forms than in Lewis's day.

are fascinating, he never succeeded—but modern logic is gradually moving closer and closer to his dream. An important point needs to be made about the universal language of which Leibniz dreamed:

Pictographic Symbols & Constructs Allowed Leibniz knew that **UL** would have to permit *pictographic* symbols and constructs based on them. When most people think about modern formal logic today, they think only of formulae that are symbolic and linguistic in nature. This thinking reflects an ignorance of the fact that plenty of work in the rational sciences makes use of diagrams.[8]

II. Universal Calculus (**UC**) of Reasoning for Rational Thought Leibniz envisioned a system for reasoning over content expressed in **UL**, and it is this system which, as Lewis (1960) explains, is the true "precursor of symbolic logic (p. 9)." Indeed, Lewis (1960) explains that there are seven "principles of [Leibniz's] calculus" (= of—for us—**UC**), and some of them are strikingly modern. For instance, according to Principle 1, that "[W]hatever is concluded in terms of certain variable letters may be concluded in terms of any other letters which satisfy the same conditions," we are allowed to make deductive inferences by instantiating schemas, a technique at the very heart of *Principia Mathematica*, and still very much alive today.[9]

III. Distinction Between Extensional and Intensional Logic Part of Leibniz's seminal work in logic is in his taking account of not only ordinary objects in the extensional realm, but "possible" objects "in the region of ideas." Leibniz even managed to make some of the key distinctions that drive the modal logic of possibility and necessity. For example, he distinguished between things that don't exist but could, versus things that don't exist and can't possibly exist; and between those things that exist necessarily versus those things that exist contingently. Clearly the influence of Leibniz on the thinker who gave us the first systems of modal logic, C.I. Lewis, was significant.[10]

IV. Welcoming Infinitary Objects and Reasoning Leibniz spent quite a bit time thinking about infinitary objects and reasoning over them. This is of course a massive understatement, in light of his invention of the calculus (see Footnote 8; and note that $\int$ was also given to us by Leibniz), which for him was based on the concept of an infinitesimal, *not* on the concept that serves today as a portal for those introduced to the differential and integral calculus: the concept of a limit.[11]

To conclude this section, we point out that while Hobbes without question advanced the notion that reasoning is computation (e.g., see Hobbes 1981), he had a very limited conception of reasoning as compared to Leibniz, and also had a relatively limited conception of computation as well.

---

[8] Lewis (1960) asserts that the universal language, if true to Leibniz, would be ideographic pure and simple. We disagree. Ideograms can be pictograms, and that possibility is, as we note here, welcome—but Leibniz also explicitly wanted to allow **UL** to allow for traditional symbolic constructions, built out of *non*-ideographic symbols. We know this because of Leibniz's seminal work in connection with giving us (the differential and integral) calculus, which is after all routinely taught today using Leibniz's symbolic notation (e.g., $\frac{dx}{dy}$, where $y = f(x)$).

[9] Axiomatic treatments of arithmetic, e.g., make use of this rule of inference, and then need only add *modus ponens* for (first-order & finitary) proof-theoretic completeness.

[10] The systems of modal logic for which C.I. Lewis is rightly famous are given in the landmark (Lewis and Langford 1932).

[11] An interesting way to see the ultimate consequences, for formal logic, of Leibniz's infinitary reasoning in connection with infinitesimals and calculus, is to consider in some detail, from the perspective of formal logic, the "vindication" of Leibniz's infinitesimal-based provided by Robinson (1996). Space limitations make the taking of this way herein beyond scope.

12.3 The Leibnizian "Three-Ray" Conception of AI

Inspired by Leibniz's vision of the "art of infallibility," in the form of both **UL** and **UC**, a heterogenous logic powerful enough to express and rigorize all of systematic human thought, we can nearly always position some particular AI work we are undertaking within a view of logic that allows a particular logical system to be positioned relative to three color-coded dimensions, which correspond to the three arrows shown in Fig. 12.1. The blue ray corresponds to purely extensional logical systems, the green the intensional logical systems, the orange to infinitary logical systems (and the red to diagrammatic logical systems). (In the interests of space, we leave aside the fact that each of the three main rays technically has a red sub-ray, which holds those logics that have pictographic elements, in addition to standard symbolic ones. We only show one red logic in Fig. 12.1.)

We have positioned our own logical system $\mathcal{DCEC}^*$, which we use to formalize a good deal of human communication and cognition (for a recent example, see e.g. Bringsjord et al. 2014) within Fig. 12.1; it's location is indicated by the black dot therein, which the reader will note is quite far down the dimension of increasing expressivity that ranges from expressive extensional logics (e.g., FOL and SOL), to logics with intensional operators for knowledge, belief, and obligation (so-called philosophical logics; for an overview, see Goble 2001). Intensional operators like

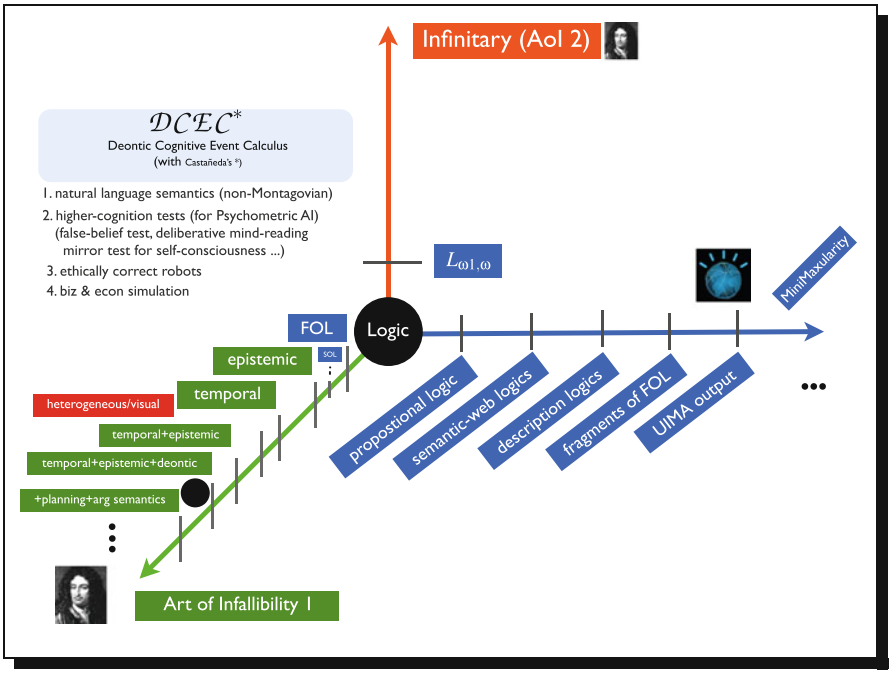


Fig. 12.1 Locating $\mathcal{DCEC}^*$ in "three-ray" Leibnizian universe

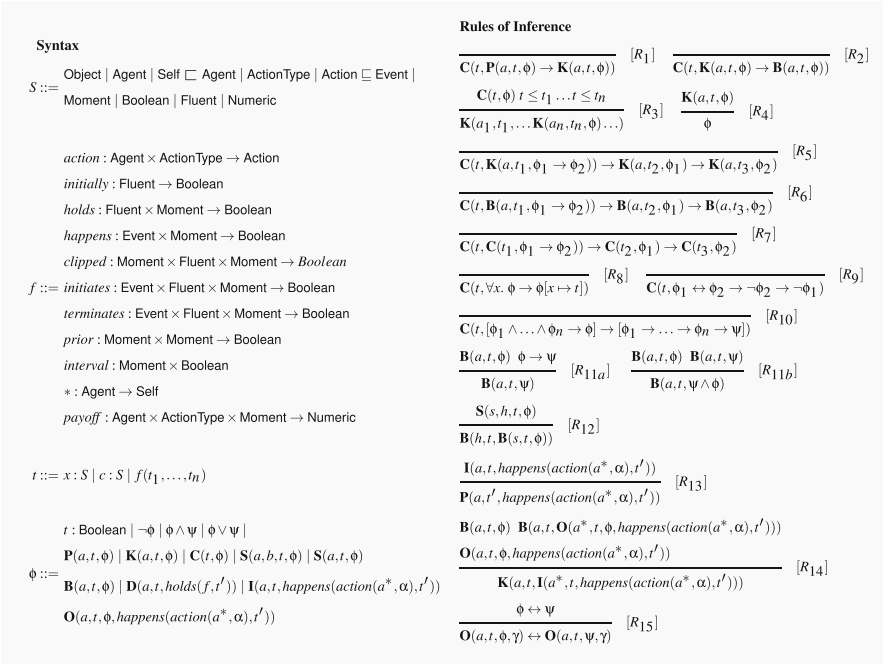


Fig. 12.2 $\mathcal{DCEC}^*$ syntax and rules of inference

these are first-class elements of the language for $\mathcal{DCEC}^*$. This language is shown in Fig. 12.2. The reader will see that the language is at least generally aligned and inspired by Leibniz's **UL** and **UC**.

What does the three-ray framework described above imply about the theory and future of AI? Assuming that every AI system or AI agent can be placed within our three-ray conceptualization, which is directly inspired by Leibniz's art of infallibility, every AI system or AI agent can be placed at a point along one of the three color-coded axes in this conceptualization. (Watson, for instance, has been placed on the blue axis or ray.)

We can quickly note that contemporary AI is most a "blue" affair, with some "green" activity.

As promised, we turn now to a framework that is more fine-grained than the three-ray one.

12.4 An "Engineeringish" Leibnizian Theory of AI

We shall be extremely lattitudinarian in our suppositions about the composition of a **logical system** $\mathscr{L}$, guided to a significant degree by the earlier framework presented in Bringsjord (2008), and by the points I–IV we have just made about Leibniz's work: Each such system will be a quintuple $\langle \mathcal{F}, \mathcal{A}, \mathcal{S}_{\mathcal{A}}, \mathcal{S}_{\mathcal{F}}, \mathcal{M} \rangle$:

Logical System $\mathscr{L}$

- $\mathcal{F}$ A formal language defined by an alphabet and a grammar that generates a space of well-formed formulae. We here use 'formula' and 'formulae' in an aggressively ecumenical way that takes account of the founding conceptions and dreams of Leibniz, so that formulas can for instance even be diagrammatic in nature. For example, in keeping with (Arkoudas and Bringsjord 2009), our relaxed concept of *formula* will allow diagrammatic depictions of seating puzzles to count as formulae. Another example pointing to how wide a scope we envisage for $\mathcal{F}$ is that any systematic kind of probabilistic or strength-factor parameters are permitted to be included in, or attached to, formulae.
- $\mathcal{A}$ An "argument theory" that regiments the concept of a structured, linked case in support of some formula as conclusion. Notice that we don't say 'proof theory.' Saying this would make the second element of a logical system much too narrow. We take proofs to be a special case of arguments.
- $\mathcal{S}_{\mathcal{A}}$ A systematic scheme for assessing the validity of a given argument α in $\mathcal{A}$.
- $\mathcal{S}_{\mathcal{F}}$ A systematic scheme for assessing the semantic value of a given formula ϕ in $\mathcal{F}$.
- $\mathcal{M}$ Finally, a metatheory consisting of informative theorems, expressed and established in classical, deductive ways, regarding the other four members of $\mathscr{L}$, their inter-relationships, and connections to other parts of the formal (theorem-driven) sciences.

There are well-known theorems in foundational computability theory that let us assert the following general claim. In the case of computer programs written for Turing machines, the logical system would be first-order logic.

Simulation in Logic
Given any computer program W that maps inputs from I to outputs in O, its processing can be captured by search for arguments in a logic system $W_{\mathcal{L}}$.

- Inputs would be be formulae from $I_{\mathcal{F}}$ whose validity would have to be established.
- Outputs $O_{\mathcal{F}}$ will be computed by filling in *slots* in the inputs if a successful argument has been found for the input.

Given that we can simulate an AI system in a logic system, we could then evaluate the AI system against Leibniz's goal for a universal logic by looking for the minimal logic needed to simulate the AI system. There are several criteria for

ranking logics, but we feel that they are not suited for evaluating AI systems from the standpoint of Leibniz's dream. We propose the following Leibnizian criteria to evaluate logics when they are used to simulate AI. The criteria can be broken up into representation and reasoning (in keeping with aspects I./**UL** and II./**UC** from Sect. 12.2.2). An earlier version of these criteria can be found in Govindarajulu et al. (2013). Very broadly, the representation criteria let us evaluate how good a system is at representing various mental and worldly objects. The reasoning criteria let us evaluate how good the system is at representing various kinds of reasoning processes. We assign numeric scores for different criteria; the higher a score, the more desirable it is.

12.4.1 Criteria for $\mathcal{F}$

We propose the following criteria to establish how strong a representation system is.

12.4.1.1 Degree of Coverage

A representation system should be broad enough to cover all possible types of signs and structures. Pierce's *Theory of Signs* provides a broad account of the types of signs that one could expect a universal logic to possess. Briefly, an *icon* has a morphological mapping to the object it denotes. A *symbol* has no such mapping to the object it denotes.[12] Mathematical logic used in AI has largely focused on symbolic formulae. There has been some work in using iconic or visual logics (e.g., see the Vivid system introduced in Arkoudas and Bringsjord 2009).) A large part of human reasoning would be difficult to reduce to a purely symbolic form. In Govindarajalulu et al. (2014), we show some examples of visual/geometric reasoning used to prove theorems in special relativity theory (Table 12.1).

Table 12.1 Degree of coverage

Coverage		
Score	Interpretation	Example
1	Only symbolic	Propositional logic, probability calculus, etc.
1	Only iconic	Image processing
2	Both iconic and symbolic	Vivid

[12]The distinction is not always perfect. E.g., {} for the empty set is both iconic and symbolic.

Table 12.2 Degree of quantification/size

Quantification/size		
Score	Interpretation	Example
0	Propositional	SAT solvers
0.5	First-order with limited quantification	OWL (Baader et al. 2007)
1	First-order	Mizar (Naumowicz and Kornilowicz 2009)
2	Second-order, higher-order logics	Reverse mathematics project (Simpson 2009)
3	Infinitary logics etc.	ω-rule (Baker et al. 1992)

12.4.1.2 Degree of Quantification/Size

This captures how many individual objects and concepts a formula can capture (Table 12.2). Terms referring to sets or classes of objects would be counted as referring to only a single object. At the lowest level in this scale, we have simple propositional systems. At the highest levels, we have infinitary logics capable of expressing infinitely long statements. Such statements would be needed to formally capture some statements in mathematics which cannot be expressed in a finite fashion (Barwise 1980).[13]

12.4.1.3 Degree of Homoiconicity

Ideally, the representation system should be able to represent its own formulae and structures in itself easily. The term "homoiconicity" is usually used to refer to the Lisp class of languages which are capable of natively representing, to various degrees, their own programs. First-order logic does not have this capability natively, but one can achieve this by using schemes like Gödel numbering. In Bringsjord and Govindarajulu (2012), we go through several such schemes for first-order logic (Table 12.3).

The earliest such scheme, due to Gödel, assigns a natural number n^{ϕ} for every formula ϕ. Then every such number n^{ϕ} is assigned a term $\hat{n}^{\phi}$ in the language. Thus one could write down sentences in the fashion $\chi(\hat{n}^{\phi})$, in which we assert χ about ϕ. A more recent scheme termed *reification* assigns terms in a first-order language $\mathcal{L}_{main}$ to formulae and terms in another first-order language $\mathcal{L}_{obj}$. The *event calculus* (nicely covered and put to AI-use in Mueller 2006) formalism in AI employs this scheme. In this scheme, states of the world are formulae in $\mathcal{L}_{obj}$ and are called

[13] Infinitary logics are "measured" in terms of length not only of formulae, but e.g. number of quantifiers permitted in formulae. In the present paper, we leave aside the specifics. But we do point out to the motivated reader that Fig. 12.1, on the infinitary "ray," refers to the infinitary logic $\mathcal{L}_{\omega_1\omega}$, which, in keeping with the standard notation, says that disjunctions/conjunctions can be of a countably infinite length, whereas the number of quantifiers allowed in given formulae must be finite.

Table 12.3 Degree of homoiconicity

Homoiconicity		
Score	Interpretation	Example
0	No native capability	First-order logic
1	Can talk about formulae only	First-order modal logic Bringsjord et al. (2013)
1	Can talk about arguments only	First-order denotational proof languages
2	Arguments and formulae	

Table 12.4 Degree of possibility

Possibility		
Score	Interpretation	Example
0	No possibility	Propositional logic
1	One mode of possibility	Markov logic (Richardson and Domingos 2006)
2	Multiple modes of possibility	First-order logic (Halpern 1990)

fluents. For example, formulae in $\mathcal{L}_{main}$ stating that the sentence *raining* in $\mathcal{L}_{obj}$ never holds would be:

$$\neg \forall t : holds(raining, t)$$

Ideally, we should be able to use $\mathcal{F}$ to state things not just about objects in $\mathcal{F}$, but about $\mathcal{S}_A$, $\mathcal{S}_F$, and $\mathcal{L}$ itself in general. For example, first-order logic cannot directly represent first-order proofs, but first-order denotational proof languages can do so (Arkoudas 2000).

12.4.1.4 Degree of Possibility

Can the formulae represent alternate states-of-affairs or yet unrealized states-of-affairs? The system should be expressive enough to represent uncertain information, information that is false but could be true under alternate states of affairs, information that could hold in the future, etc.; notice the connection here to our point III about Leibniz's art of infallibility. Note that by including possibility, we also include probablility and uncertainty (Table 12.4).

12.4.1.5 Degree of Intensionality

How good is the system at representing other agents' mental states? The system should be able to represent and reason over knowledge, belief, and other intensionalities of other agents and itself. While one could refurbish a purely extensional system to represent intensionalities by various "squeezing" schemes, such squeezing leads to problems. In Bringsjord and Govindarajulu (2012), we look

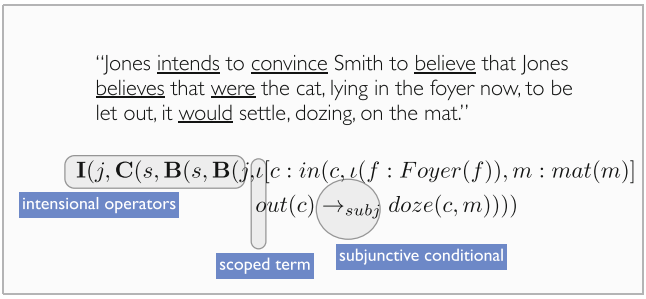


Fig. 12.3 Representing intensionality

Table 12.5 Degree of intensionality

Intensionality		
Score	Interpretation	Example
0	No intensionality	First-order logic
1	Restricted intensionality	Modal logic with no iterated beliefs
2	Full intensionality	Unlimited iterated beliefs Bringsjord et al. (2013)

at various schemes for squeezing *knowledge* into first-order logic and show how each of the schemes is either incoherent or leads to outright contradictions. These results suggest that a universal logic should have the facility to directly represent intensionalities. What would such a scheme look like? A complex intensional statement and its possible representation in a semi-formal first-order modal notation is shown in Fig. 12.3 (Table 12.5).

12.4.2 Criteria for $\mathcal{S}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{F}}$

$\mathcal{S}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{F}}$ capture reasoning in $\mathcal{L}$. We would want them to be as general as possible. For measuring generality, we borrow a yardstick from computability theory. Since $\mathcal{S}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{F}}$ are computer programs, we can measure how general they are by looking at where they are placed in the computability heirarchy. In this heirarchy, the lower levels would be populated by the *computable* schemes. Within the computable class, we would have the programs ordered by their asymptotic complexity. After the computable class, we would have the *semi-computable* class of programs. The justification is that the higher a scheme is placed, the greater the number of kinds of reasoning processes it would be able to simulate (Table 12.6).

Table 12.6 Criteria $\mathcal{S}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{F}}$

$\mathcal{S}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{F}}$		
Score	Interpretation	Example
0	Computable and tractable	Nearest neighbors
1	Computable and intractable	Bayesian networks (Chickering 1996)
2	Uncomputable	First-order theoremhood (Boolos et al. 2007)

Table 12.7 Criteria $\mathcal{M}$

Boundary theorems		
Score	Interpretation	Example
0	No boundary theorems	Adhoc learning systems
1	Soundness or completeness	AIXI no notion of soundness (Hutter 2005)
2	Soundness and completeness	First-order logic (Boolos et al. 2007)

12.4.3 Criteria for $\mathcal{M}$

Logical systems are accompanied by a slew of meta-theorems. These frequently are soundness and completeness results, to start. Logical systems usually have other meta-theorems that might not be, at least on the surface, that useful for evaluating AI systems; for example, the compactness theorem for first-order logic. We require that the logic system contain, at the least, two theorems we term the *boundary theorems*. Boundary theorems correspond roughly to soundness and completeness (Table 12.7).

12.5 Watson Abstractly

We now look at a highly abstract model of how Watson works.[14] The abstract model is what an ideal-observer logicist would see when examining Watson. Looking at Watson's initial stages in Fig. 12.4, we see that for a given question Q, multiple answers $\{\mathsf{A}_1, \mathsf{A}_2, \ldots, \mathsf{A}_n\}$ are generated. Each of these answers is paired with multiple pieces of data, the *evidence*, resulting in $\langle e_1^{\mathsf{A}_i}, \ldots, e_m^{\mathsf{A}_i} \rangle$. Each such tuple $\langle \mathsf{Q}; \mathsf{A}_i; e_1^{\mathsf{A}_i}, \ldots, e_m^{\mathsf{A}_i} \rangle$ is then fed through *answer and evidence scorers*, resulting in a *feature vector* $v_{\langle \mathsf{Q}, \mathsf{A}_i \rangle}$ for each $\langle \mathsf{Q}, \mathsf{A}_i \rangle$ pair.

Each feature vector is then passed through multiple *phases* as shown in Fig. 12.5. Within each phase, there are classifiers for different broad categories of questions

[14] A more concrete overview of Watson can be found in Ferrucci et al. (2010). While (Ferrucci and Murdock 2012) delves much more into Watson, for the material in this section one needs to primarily refer to Gondek et al. (2012).

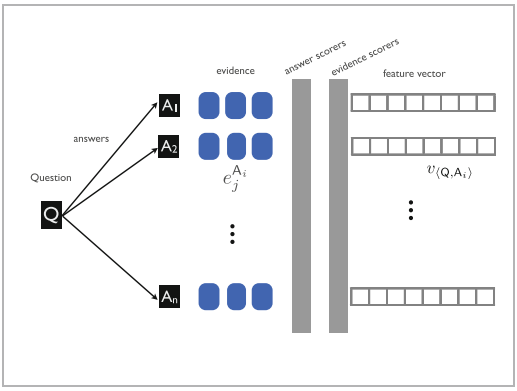


Fig. 12.4 Initial stages Watson pipeline

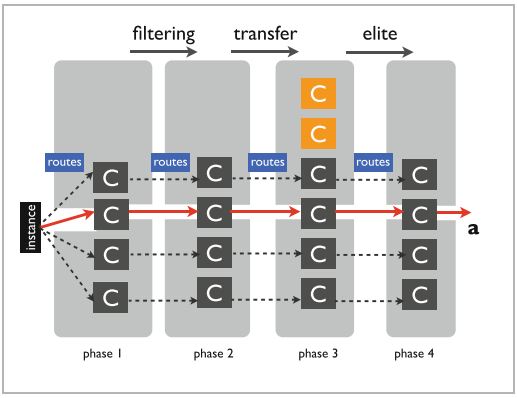


Fig. 12.5 Final stages

(e.g., Date, Number, Multiple Choice etc.). Any given $v_{\langle Q,A_i\rangle}$ passes through at most one classifier in a phase. The phases essentially serve to reduce the vast number of answers, so that a final ranking can focus only on an elite set of answers. Each classifier performs logistic regression on the input, as shown in Fig. 12.6. The classifier maps the feature vector into a confidence score $c \in [0, 1]$.

While the details of the phases are not important for our discussion, we quickly sketch an abstract version of the phases. For any given question Q, all the $v_{\langle Q,A_i\rangle}$ are fed into one classifier in an initial *filtering* phase. This reduces the set of answers from n to a much smaller n', the top n' based on scores from the classifer. Then in a *normalization* phase, the feature values are first normalized with respect to the n' instances. The answers are then re-ranked by one classifier in this phase. The third *transfer* phase has specialized classifers to account for rare categories (e.g. Etymology). The final *elite* phase then outputs a final ranking of a smaller set of answers from the previous phase.

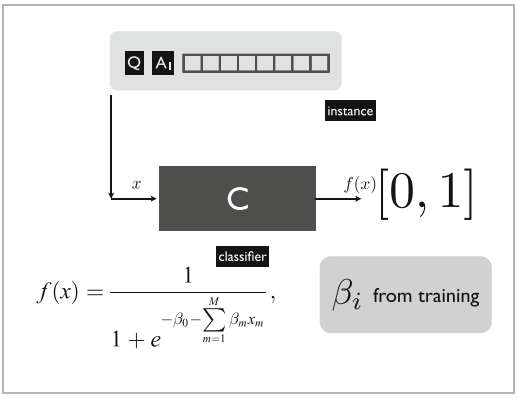


Fig. 12.6 Assigning confidence scores

12.5.1 Locating Watson

We confess that there is some "squeezing" that must be done in order to place Waston within the three-ray space. We accomplish this squeezing by again invoking an "ideal-observer" perspective when looking at Watson. This perspective is generated by imagining that an ideal observer, possessed of great intellectual powers, and able to observe the entire pipeline that takes a question Q and returns an answer A, offers a classical argument in support of A being the correct answer. Our ideal observer is well-versed in all manner of statistical inference and machine learning, and has at his command not only what happens internally when Q as input yields A as output, but also fully comprehends all the prior processing that went into the tuning of the pipeline that in the case at hand gave A from Q. From this perspective, we obtain the following:

Watson			
Ray	Criterion	Score	Reason
Ray 1	Coverage	1	Only symbolic at present
	Quantification	0.5	First-order (some first-order components)
	Homoiconicity	1	First-order
	Possibility	1	Support for probability
	Intensionality	0	No support for intensionalities
Ray 2	$\mathcal{S}_{\mathcal{A}}$ and $\mathcal{S}_{\mathcal{F}}$	1	Sub-first-order reasoning
Ray 3	Boundary theorems	0	No formal study of Watson yet

12.6 The Future of AI

We end with some brief remarks about the future of AI in light of the foregoing.

These days there's much talk about the so-called "Singularity," the point in time at which AI systems, having reached the level of human intelligence, create AI^{+}

systems that *exceed* this level of intelligence—which then leads to an explosion in which AI^{+} systems build even smarter systems, and the process (AI^{++}, AI^{+++}, ...) iterates, leaving the poor humans in the dust.[15] Given our Leibnizian analysis above, anything like the Singularity looks, well, a bit ... unreasonably sanguine. In terms of the above three-ray space, the reason is that, one, AI is currently *defined* as a "blue" enterprise, with just a tinge of "green" now in play. Can machines in the blue range create machines that range across the green, red, and orange axes? No. And this negative is a matter of mathematical fact. For just as no finite-state automaton can create a full-fledged Turing machine, no blue logic can create, say, an orange one. So *humans* must carry AI from blue into the the other colors. And the fact is, there just is nothing out there about how that lifting is going to work, if it indeed can. On the other hand, Watson tells us something that may provide *some* basis for the kind of optimism that "true believers" radiate these days. The something to which we allude is simply the stark fact that although the scorecard earned above by Watson, within the larger Leibnizian context we have presented, shows that it's fundamentally severely limited, the fact is that it's *behavior* in rigidly controlled environments is surprisingly impressive. It remains to be seen is how much of what humans do for a living would be replaceable by such mechanical behavior.

References

Arkoudas, K. (2000). *Denotational Proof Languages*. PhD thesis, MIT.

Arkoudas, K., & Bringsjord, S. (2009). Vivid: An AI framework for heterogeneous problem solving. *Artificial Intelligence, 173*(15), 1367–1405. http://kryten.mm.rpi.edu/vivid_030205.pdf, the url http://kryten.mm.rpi.edu/vivid/vivid.pdf provides a preprint of the penultimate draft only. If for some reason it is not working, please contact either author directly by email.

Baader, F., Calvanese, D., & McGuinness, D. (Eds.). (2007). *The description logic handbook: Theory, implementation* (2nd ed.). Cambridge: Cambridge University Press.

Baker, S., Ireland, A., & Smaill, A. (1992). On the use of the constructive omega-rule within automated deduction. In *Logic Programming and Automated Reasoning* (pp. 214–225). Berlin/New York: Springer.

Barwise, J. (1980). Infinitary logics. In E. Agazzi (Ed.), *Modern logic: A survey* (pp. 93–112). Dordrecht, Reidel.

Boolos, G. S., Burgess, J. P., & Jeffrey, R. C. (2007). *Computability and logic* (5th ed.). Cambridge: Cambridge University Press.

Bringsjord, S. (1992). *What robots can and can't be*. Dordrecht: Kluwer.

Bringsjord, S. (1998). Chess is too easy. *Technology Review, 101*(2), 23–28. http://www.mm.rpi.edu/SELPAP/CHESSEASY/chessistooeasy.pdf

[15]The primogenitor of the case for this series of events is Good (1965). A modern defense of the this original case is provided by Chalmers (2010). A formal refutation of the original and modernized argument is supplied by Bringsjord (2012). An explanation of why belief in the Singularity is fideistic is provided by Bringsjord et al. (2013).

Bringsjord, S. (2008). Declarative/logic-based cognitive modeling. In R. Sun (Ed.), *The handbook of computational psychology* (pp. 127–169). Cambridge: Cambridge University Press. http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf.

Bringsjord, S. (2011). Psychometric artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence, 23*(3), 271–277.

Bringsjord, S. (2012). Belief in the singularity is logically Brittle. *Journal of Consciousness Studies, 19*(7), 14–20. http://kryten.mm.rpi.edu/SB_singularity_math_final.pdf

Bringsjord, S., & Arkoudas, K. (2006). On the provability, veracity, and AI-relevance of the church-turing thesis. In A. Olszewski, J. Wolenski, & R. Janusz (Eds.), *Church's thesis after 70 years* (pp. 66–118). Frankfurt: Ontos Verlag. http://kryten.mm.rpi.edu/ct_bringsjord_arkoudas_final.pdf, This book is in the series *Mathematical Logic*, edited by W. Pohlers, T. Scanlon, E. Schimmerling, R. Schindler, and H. Schwichtenberg.

Bringsjord, S., & Ferrucci, D. (2000). *Artificial intelligence and literary creativity: Inside the mind of brutus, a storytelling machine*. Mahwah: Lawrence Erlbaum.

Bringsjord, S., & Govindarajulu, N. S. (2012). Given the web, what is intelligence, really? *Metaphilosophy, 43*(4), 361–532. http://kryten.mm.rpi.edu/SB\NSG\Real\Intelligence\040912.pdf, This URL is to a preprint of the paper.

Bringsjord, S., & Govindarajulu, N. S. (2013). Toward a modern geography of minds, machines, and math. In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence* (Studies in applied philosophy, epistemology and rational ethics, Vol. 5, pp. 151–165). New York: Springer. doi:10.1007/978-3-642-31674-6_11, http://www.springerlink.com/content/hg712w4l23523xw5

Bringsjord, S., & Licato, J. (2012). Psychometric artificial general intelligence: The Piaget-MacGuyver room. In P. Wang & B. Goertzel (Eds.), *Foundations of artificial general intelligence* (pp. 25–47). Amsterdam: Atlantis Press. http://kryten.mm.rpi.edu/Bringsjord_Licato_PAGI_071512.pdf, This url is to a preprint only.

Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI–03)* (pp. 887–893). San Francisco: Morgan Kaufmann. http://kryten.mm.rpi.edu/scb.bs.pai.ijcai03.pdf

Bringsjord, S., Bringsjord, A., & Bello, P. (2013). Belief in the singularity is fideistic. In A. Eden, J. Moor, J. Søraker, & E. Steinhart (Eds.), *The singularity hypothesis* (pp. 395–408). New York: Springer.

Bringsjord, S., Govindarajulu, N., Ellis, S., McCarty, E., & Licato, J. (2014). Nuclear deterrence and the logic of deliberative mindreading. *Cognitive Systems Research, 28*, 20–43.

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*, 7–65.

Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In D. Fisher & H.-J. Lenz (Eds.), *Learning from data* (Lecture Notes in Statistics, Vol. 112, pp. 121–130). Berlin: Springer.

Ferrucci, D. A., & Murdock, J. W. (Eds.). (2012). *IBM Journal of Research and Development, 56*. http://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6177717

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine, 31*, 59–79. http://www.stanford.edu/class/cs124/AIMagzine-DeepQA.pdf

Goble, L. (Ed.). (2001). *The Blackwell guide to philosophical logic*. Oxford: Blackwell Publishing.

Gondek, D., Lally, A., Kalyanpur, A., Murdock, J. W., Dubou, P. A., Zhang, L., Pan, Y., Qiu, Z., & Welty, C. (2012). A framework for merging and ranking of answers in DeepQA. *IBM Journal of Research and Development, 56*(3.4), 14:1–14:12.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machines. In F. Alt & M. Rubinoff (Eds.), *Advances in computing* (Vol. 6, pp. 31–38). New York: Academic.

Govindarajalulu, N., Bringsjord, S., & Taylor, J. (2014). Proof verification and proof discovery for relativity. *Synthese*, 1–18. doi:10.1007/s11229-014-0424-3, http://dx.doi.org/10.1007/s11229-014-0424-3

Govindarajulu, N. S., Bringsjord, S., & Licato, J. (2013). On deep computational formalization of natural language. In M. H. A. Abdel-Fattah & K. -U. Kühnberger (Ed.), *Proceedings of the Workshop: "Formalizing Mechanisms for Artificial General Intelligence and Cognition" (Formal MAGiC) at Artificial General Intelligence 2013*. http://cogsci.uni-osnabrueck.de/~formalmagic/FormalMAGiC-Proceedings.pdf

Halpern, J. (1990). An analysis of first-order logics of probability. *Artificial Intelligence, 46*(3), 311–350.

Hobbes, T. (1981). Computatio sive logica: logic De corpore, Part 1. Abaris, Norwalk, CT, This volume provides both an English and Latin version of Part 1 (Chaps. 1–6) of Hobbes's book—as it's known—*De Corpore*, which was first published in 1655. The editors are: I. Hungerland (Editor), G. Vick; translator: A. Martinich.

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. New York: Springer.

Leibniz, G. W. (1991). Discourse on metaphysics • correspondence with Arnauld • monadology. Open court, LaSalle, IL, this is the twelfth printing. First published in 1902 (G. Montgomery, Trans.; translation modified by Albert Chandler).

Leibniz, G., & Gerhardt, C. (1890). *Philosophischen Schriften von Leibniz* (Vol. 7). Berlin: Weidmann. DE, gerhardt is the editor.

Lewis, C. I. (1960). *A survey of symbolic logic: The classic algebra of logic*. New York: Dover.

Lewis, C. I., & Langford, C. H. (1932). *Symbolic logic*. New York: Century Company.

Mueller, E. (2006). *Commonsense reasoning*. San Francisco: Morgan Kaufmann.

Naumowicz, A., & Kornilowicz, A. (2009). A brief overview of Mizar. DOI Retrieved on July 26, 2013. http://dx.doi.org/10.1007/978-3-642-03359-9_5

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning, 62*(1–2), 107–136.

Robinson, A. (1996). *Non-standard analysis*. Princeton: Princeton University Press. This is a reprint of the revised 1974 edition of the book. The original publication year of this seminal work was 1966.

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River: Prentice Hall.

Simpson, S. G. (2009). *Subsystems of second order arithmetic* (Vol. 1). Cambridge/New York: Cambridge University Press.

Chapter 13
The Computational Theory of Cognition

Gualtiero Piccinini

Abstract According to the computational theory of cognition (CTC), cognitive capacities are explained by inner computations, which in biological organisms are realized in the brain. Computational explanation is so popular and entrenched that it's common for scientists and philosophers to assume CTC without argument. But if we presuppose that neural processes are computations before investigating, we turn CTC into dogma. If, instead, our theory is to be genuinely empirical and explanatory, it needs to be empirically testable. To bring empirical evidence to bear on CTC, we need an appropriate notion of computation. In order to ground an empirical theory of cognition, as CTC was designed to be, a satisfactory notion of computation should satisfy at least two requirements: it should employ a robust notion of computation, such that there is a fact of the matter as to which computations are performed by which systems, and it should not be empirically vacuous, as it would be if CTC could be established a priori. In order to satisfy these requirements, the computational theory of cognition should be grounded in a mechanistic account of computation. Once that is done, I evaluate the computational theory of cognition on empirical grounds in light of our best neuroscience. I reach two main conclusions: cognitive capacities are explained by the processing of spike trains by neuronal populations, and the processing of spike trains is a kind of computation that is interestingly different from both digital computation and analog computation.

Keywords Computational theory of cognition • Neural computation • Explanation • Mechanism • Spike trains • Digital computation • Analog computation

---

This paper is a substantially revised and updated descendant of Piccinini 2007, which it supersedes. Accounts of computation in the same spirit are also defended in Fresco 2014 and Milkoswki 2013. Thanks to an anonymous referee for helpful comments. Thanks to Elliott Risch for editorial assistance. This material is based on work supported in part by a University of Missouri research award.

G. Piccinini (✉)
University of Missouri-St. Louis, 1 University Blvd, St. Louis, MO, USA
e-mail: piccininig@umsl.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_13

13.1 Cognitive Capacities

When we explain the capacities of certain physical systems, we appeal to the computations they perform. For example, calculators—unlike, say, air conditioners—have the peculiar capacity of performing multiplications: if we press appropriate buttons on a (well-functioning) calculator in the appropriate order, the calculator yields the product of the input data. Our most immediate explanation for this capacity is that, under the relevant conditions, calculators perform an appropriate computation—a multiplication—on the input data. This is a paradigmatic example of computational explanation.

Animals exhibit *cognitive capacities*: capacities to respond to their environments in extraordinarily subtle, specialized, and adaptive ways—ways that require them to acquire and process information about their environments. In explaining those capacities, we often appeal to cognitive states such as perceptions, memories, intentions, etc. We also recognize that the mechanisms that underlie cognitive capacities are *neural*—no nervous systems, no cognitive systems.[1] But it is difficult to connect cognitive constructs to their neural realizers—to see how perceptions, memories, intentions, and the like, are realized by neural states and processes. In various forms, this problem has haunted the sciences of cognition since their origin.

In the mid-twentieth century, Warren McCulloch and others devised an ingenious solution: cognitive capacities are explained by computations realized in the brain (McCulloch and Pitts 1943; Wiener 1948; von Neumann 1958; Piccinini 2004a provides some background). This is the computational theory of cognition (CTC), which explains cognitive capacities more or less in the way we explain the capacities of computing systems. There are many versions of CTC: *classical* versions, which tend to downplay what we know about neural systems (e.g., Newell and Simon 1976; Fodor 1975; Pylyshyn 1984; Gallistel and King 2009); *connectionist* versions, which pay lip service to it (e.g., Rumelhart and McClelland 1986); and *neurocomputational* versions, which are more or less grounded in what we know about neural systems (e.g., Churchland and Sejnowski 1992; O'Reilly and Munakata 2000). According to all of them, the nervous system is a computing system, and some of its capacities—including its cognitive capacities—are explained by its computations.

CTC has encountered resistance. Some neuroscientists are skeptical that the brain may be adequately characterized as a computing system in the relevant sense (e.g., Gerard 1951; Rubel 1985; Perkel 1990; Edelman 1992; Globus 1992). Some psychologists think computational explanation of cognitive capacities is inadequate (e.g., Gibson 1979; Varela et al. 1991; Thelen and Smith 1994; Port and van Gelder 1995; Johnson and Erneling 1997; Ó Nualláin et al. 1997; Erneling and

[1] Some philosophers have argued that the realizers of cognitive states and processes include not only the nervous system but also some things outside it (e.g., Wilson 2004). I will ignore this possible complication because this simplifies the exposition without affecting my conclusions.

Johnson 2005). And some philosophers find it implausible that everything about cognition may be explained by computation (e.g., Taube 1961; Block 1978; Putnam 1988; Maudlin 1989; Mellor 1989; Searle 1992; Bringsjord 1995; Dreyfus 1998; Harnad 1996; Penrose 1994; van Gelder 1995; Wright 1995; Horst 1996; Lucas 1996; Copeland 2000; Fetzer 2001). Whether CTC explains every aspect of every cognitive capacity remains controversial. But without a doubt, CTC is a compelling theory.

Digital computers are more similar to cognitive systems than anything else known to us. Computers can process information, perform calculations and inferences, and exhibit a dazzling variety of cognitive capacities, including that of guiding sophisticated robots. Computers and natural cognitive systems are sufficiently analogous that CTC is attractive to most of those who are searching for a mechanistic explanation of cognition. As a consequence, CTC has become the mainstream explanatory framework in psychology, neuroscience, and naturalistically inclined philosophy of cognitive science. In some quarters, CTC is now so entrenched that it seems commonsensical.

More than half a century after CTC's introduction, it's all too easy—and all too common—to take for granted that cognitive capacities are explained by neural computations. But if we *presuppose* that neural processes are computations before investigating, we turn CTC into dogma. For our theory to be genuinely empirical and explanatory, it needs to be empirically testable. To bring empirical evidence to bear on CTC, we need an appropriate notion of computation and a corresponding notion of computational explanation.

In order to ground an empirical theory of cognition, as CTC was designed to be, a satisfactory notion of computation should satisfy at least two requirements. First, it should be robust in the sense that there is a fact of the matter as to which computations are performed by which systems. This might be called the *robustness requirement*. Second, it should not be empirically vacuous, as it would be if CTC could be established a priori. This might be called the *non-vacuity requirement*. This paper explicates the notion of computation so as to satisfy these requirements and discusses whether it plausibly applies to the explanation of cognition.

13.2 Computation and Representation

According to the popular *semantic account* of computation, a computation is a process that manipulates representations in an appropriate way (e.g., Cummins 1983; Churchland and Sejnowski 1992 Fodor 1998; Shagrir 2001, 2006). According to the typical semantic account, computations are individuated at least in part by their semantic properties. The semantic account is appealing because it fits well both our practice of treating the internal states of computing systems as representations and the representational character of those cognitivist constructs, such as perceptions and intentions, that are traditionally employed in explaining cognitive capacities.

While the semantic account is correct to emphasize the importance of representation in cognition, it is inadequate as an account of computation per se. Here I have room only for a few quick remarks; I have discussed the semantic view in detail elsewhere (Piccinini 2004b).

For present purposes, we need to distinguish between what may be called essential representations and accidental ones. Essential representations are individuated, at least in part, by their content. In this sense, if two items represent different things, they are different kinds of representation. For instance, at least in ordinary parlance, cognitivist constructs are typically individuated, at least in part, by their content: the concept of smoke is individuated by the fact that it represents smoke, while the concept of fire is individuated by the fact that it represents fire.[2] Representations in this sense of the term have their content essentially: you can't change their content without changing what concepts they are.

By contrast, accidental representations can be individuated independently of their content; they represent one thing or another (or nothing at all) depending on whether they are interpreted and how. Strings of letters of the English alphabet are representations of this kind: they are individuated by the letters that form them, regardless of what the strings mean or even whether they mean anything at all. For instance, the string "bello" means (*of*) *war* to speakers of Latin, *beautiful* to speakers of Italian, and nothing in particular to speakers of most other languages.[3]

The semantic account of computation requires essential representations, but all it has available is accidental ones. If we try to individuate computations by appealing to the semantic properties of accidental representations, we obtain an inadequate notion of computation. For the same accidental representation may represent different things (including nothing at all) to different interpreters. As a consequence, a putative computation that is individuated by reference to the semantic properties of accidental representations may be taken by different interpreters to compute different things without changing anything in the process itself.

Just as speakers of different languages can interpret the same string of letters in different ways, under the semantic account (plus the notion of accidental representation) different observers could look at the same activity of the same mechanism and interpret it as two different computations. But a process that

[2] I am using "concept" in a pre-theoretical sense. Of course, there may be ways of individuating concepts independently of their content, ways that may be accessible to those who possess a scientific theory of concepts but not to ordinary speakers.

[3] The distinction between essential and accidental representation is closely related to the distinction between original and derived intentionality. Derived intentionality is intentionality conferred on something by something that already has it; original intentionality is intentionality that is not derived (Haugeland 1997). If something has original intentionality, presumably it is an essential representation (it has its content essentially); if something has derived intentionality, presumably it is an accidental representation. These distinctions should not be confused with the distinction between intrinsic and extrinsic intentionality. Intrinsic intentionality is the intentionality of entities that are intentional regardless of their relations with anything else (Searle 1983). Something may be an essential representation without having intrinsic intentionality, because its intentionality may be due to the relations it bears to other things.

changes identity simply by changing its observer is not a good foundation for the science of computation, let alone cognition. Such a notion of computation fails to satisfy the robustness requirement. To obtain a genuinely explanatory notion of computation, the semantic account requires the first notion of representation—essential representation. In fact, many who have explicitly endorsed the semantic account have done so on the basis of the notion of essential representation (Cf. Burge 1986; Segal 1991).

But there is no reason to believe that computational states, inputs, and outputs have their semantic properties essentially (i.e., that they are essential representations; cf. Egan 1995). On the contrary, a careful look at how computation is understood by computer scientists reveals that (digital) computational states, inputs, and outputs are individuated by the strings of letters that constitute them and computations are individuated by the operations performed on those strings—regardless of which, if any, interpretation is applied to the strings.

Psychologists and neuroscientists rarely distinguish between explanation that appeals to computation and explanation that appeals to representation; this has convinced many—especially philosophers of cognitive science—that computational explanation is essentially representational. But this is a mistake. Computational explanation appeals to inner computations, and computations are individuated independently of their semantic properties. Whether computational states represent anything, and what they represent, is another matter.

The point is *not* that *cognitive* states are *not* representations, or that if they are representations, they must be accidental ones. The point is also *not* that representations play no explanatory role within a theory of cognition; they do play such a role (see below). The point is simply that if cognitive or neural states are computational, they are not so in virtue of their semantic properties. For this reason, among others, the semantic account of computation is inadequate.

13.3 Computational Modeling

Another popular view is that a system is computational just in case a computational model describes the capacities of a system. I will refer to this as the *mapping account* of computation. According to the mapping account, roughly speaking, anything that is described by a computation is also a computing system that performs that computation (e.g., Putnam 1967; Churchland and Sejnowski 1992; Chalmers 2011). The mapping account is tempting because it appears to gain support from the widespread use of computational models in the sciences of cognition. Nevertheless, it is even less satisfactory than the semantic account.

The main difficulty with the mapping account is that it turns so many things into computing systems that it fails the non-vacuity requirement. Paradigmatic computational explanations are used to explain peculiar capacities of peculiar systems. We normally use them to explain what calculators and computers do, but not to explain the capacities of most other systems around us. When we explain

the capacities of air conditioners, lungs, and other physical systems, we employ many concepts, but we normally do not appeal to computations. This gives rise to the widespread sense that, among physical systems, only a few special ones are computing systems in an interesting sense. This is an important motivation behind CTC. The idea is that the *cognitive* capacities that organisms exhibit—as opposed to their capacities to breathe, digest, or circulate blood—may be explained by appeal to neural computations.

And yet, it is perfectly possible to build computational models of many physical processes, including respiration, digestion, or galaxy formation. According to the mapping account, this is sufficient to turn lungs, stomachs, and galaxies into computing systems, in the same sense in which calculators are computing systems and brains may or may not be. As a consequence, many things—perhaps all things—are turned into computing systems. Some authors have accepted this consequence; they maintain that everything is a computing system (e.g., Putnam 1967; Churchland and Sejnowski 1992; Chalmers 2011). As we have seen, this is counterintuitive at best, because it conflicts with our restricted use of computational explanation. Still, intuitions are often overridden by strong arguments. Why not accept that everything is a computing system?

The problem is that if everything is a computing system (in the relevant sense), computational descriptions lose their explanatory character to the point that CTC is trivialized. To determine with some precision which class of systems can be described computationally to which degree of accuracy is difficult. Nevertheless, it is obvious that everything, including many neural systems, can be described computationally with *some* accuracy. This fact, in conjunction with the mapping account of computational explanation, establishes the truth of CTC a priori, without requiring empirical investigation. But CTC was intended to be an empirical theory, grounded on an empirical hypothesis about the kinds of processes that explain cognitive capacities. If the versatility of computational description is sufficient to establish that neural systems perform computations, then this cannot be an empirical hypothesis about the explanation of cognitive capacities—it is merely the trivial application to brains of a general thesis that applies to everything. In other words, the mapping account of computational explanation renders CTC empirically vacuous.

A further complication facing the mapping account is that the same physical system may be given many computational descriptions that are different in nontrivial respects, for instance because they employ different computational formalisms, different assumptions about the system, or different amounts of computational resources. This makes the answer to the question of which computation is performed by a system indeterminate. In other words, given the mapping account, not only is everything computational, but everything also performs as many computations as it has computational descriptions. As a consequence, this notion of computational explanation fails the robustness requirement. Like the semantic account, the mapping account is inadequate for assessing CTC.

A variant of the mapping account is that a computational explanation is one that appeals to the generation of outputs on the grounds of inputs and internal states. But without some constraints on what counts as inputs and outputs of the appropriate

kind, this variant faces the same problem. Every capacity and behavior of every system can be interpreted as the generation of an output from an input and internal states. Our question may be reformulated in the following way: *which* input–output processes, among the many exhibited by physical systems, deserve to be called computational in the relevant sense?

13.4 Functional Mechanisms

To make progress on this question, we should begin by looking into the explanatory strategies employed in physiology and engineering. For brains and computers are, respectively, biological systems and complex artifacts; it is plausible that they can be understood by the same strategies that have proven successful for other biological systems and complex artifacts.

The capacities of biological systems and complex artifacts are explained mechanistically (Bechtel and Richardson 1993; Machamer et al. 2000; Glennan 2002; Craver 2007; Bechtel 2008). A mechanistic explanation involves the partition of a system into working components, the assignment of capacities to those components, and the identification of organizational relations between the components. For any capacity of a mechanism, a mechanistic explanation invokes appropriate capacities of appropriate components of the mechanism, which, when appropriately organized and exercised, constitute the capacity to be explained. The components' capacities may be explained by the same strategy, namely, in terms of the *components'* components, capacities, and organization.

For example, the mechanism of rock fragmentation by explosives involves rocks with explosives inside them. Explosions create high-pressure gases; the gases produce a strain wave in the rock; and the strain wave cracks the rock into fragments that are propelled into ballistic trajectories. There are components (rocks, explosives, gases), organizational relations (explosives and then gases *within* rocks), and capacities of the components (fragmentation capacity of the rocks, explosive capacity of the components, high pressure of the gases). When the appropriately organized components exercise their capacities, the exercise of their capacities constitutes the explanandum phenomenon (rock fragmentation by explosives).

There is an important difference between biological systems as well as artifacts and non-biological mechanisms such as rock fragmentation by explosives. In the non-biological, non-artifact cases, things just happen the way they happen—nothing can go wrong (unless there are people who have goals pertaining to the phenomena, as is often the case with rock fragmentation). By contrast, artifacts and the traits of biological systems have *teleological* functions—functions that they are *supposed to* fulfill. If they don't fulfill them, they *malfunction*.

Some philosophers of science reject teleological functions because teleological functions seem to have a mysterious normativity that doesn't fit well with the descriptive nature of science. But teleological functions are commonly invoked in both science and everyday language; they are useful in distinguishing biological

systems and artifacts from other mechanisms. If we can find a naturalistic account of the normativity of teleological function, we can retain teleological functions in our theoretical vocabulary and use them in our account of computation.

I have argued elsewhere that there is no more mystery about the normativity of teleological functions than there is mystery about the nature of life. Life is a set of natural processes whereby some complex physical systems manage to survive for some time, reproduce, and evolve over generations. If such systems ceased to survive and have enough viable offspring, they would all cease to exist. Therefore, it is essential to living beings that they survive and have some degree of inclusive fitness. In this limited sense, we may call survival and inclusive fitness *objective goals* of organisms. Any stable contribution to the goals of organisms that is performed by one of their traits is a teleological function; such functions ought to be fulfilled on pain of extinction. This account of teleological function (here barely sketched) can be generalized to artifacts, whose functions are their stable contributions to the survival and inclusive fitness of organisms. Finally, this account can be generalized to the *subjective goals*—such as pleasure, knowledge, and beauty—of organisms that are sophisticated enough to go beyond their objective goals. A teleological function is any stable contribution to a goal of an organism by either a biological trait of the organism or an artifact (Maley and Piccinini forthcoming).

For example, the capacity of a car to run is mechanistically explained by the following: the car contains an engine, wheels, etc.; under normal conditions, the engine generates motive power, the power is transmitted to the wheels, and the wheels are connected to the rest of the car so as to carry it for the ride. Generating motive power and transmitting it to the wheels are teleological functions of the car's components; if they are not fulfilled under relevant circumstances, the car is malfunctioning. Given that the capacities of biological systems and complex artifacts are explained mechanistically, it remains to be seen how mechanistic explanation relates to computational explanation.

13.5 Computational Explanation, Functional Analysis, and Mechanistic Explanation

There used to be a tendency, in the classic literature on the philosophy of cognitive science, to assimilate mechanistic explanation to computational explanation or, equivalently, to assimilate mechanistic processes to computations (see Piccinini 2004b for some background). How so? A mechanistic process is a manifestation of a set of operations organized in a certain way. Such organized operations may be represented by a flowchart or computer program. Therefore, a mechanistic process is a computation and a mechanistic explanation is computational. Or so it may seem. Something like this line of thinking is behind the popular view that everything is computational, and that some computational formalism, such as Turing machines, captures everything that can be done mechanistically.

A closely related line of thinking attempts to draw a distinction between mechanistic explanation and functional analysis (Fodor 1968; Cummins 2000). Mechanistic explanation is viewed as dealing with concrete components of a system, whereas functional analysis is viewed as dealing with the capacities, activities, or operations of such components. According to this view, functional analysis describes the activities or operations that are responsible for a phenomenon, and it does this autonomously from mechanistic explanation. As pointed out before, a series of operations can be described by a flowchart or computer program. Therefore, functional analysis is computational, and everything performs computations.

An ally of these views is the mapping account of computation. As we have seen, the mapping account attributes computations to any physical system that can be described computationally. As a result, many authors do not explicitly distinguish between computational explanation and functional analysis, or between computational explanation and mechanistic explanation (e.g., Fodor 1968; Dennett 1978; Marr 1982).

Identifying computational explanation with functional analysis or mechanistic explanation (or both) may seem advantageous because it appears to reconcile computational explanation with the well-established explanatory strategy that is in place in biology and engineering. To be sure, neuroscientists, psychologists, and computer scientists explain the capacities of brains and computers by appealing to internal states and processes. Nevertheless, a simple identification of computational explanation and functional explanation is based on an impoverished understanding of both explanatory strategies. We have already seen above that computational explanation must be more than the appeal to inputs and internal states and processes, on pain of losing its specificity to a special class of mechanisms and trivializing CTC. For an explanation to be genuinely computational, some constraints need to be put on the nature of the inputs, outputs, and internal states and processes.

A related point applies to mechanistic explanation. The strength of mechanistic explanation derives from the different kinds of concrete components and processes it appeals to. These processes are as different as digestion, refrigeration, and illumination. Of course, we *could* abstract away from the differences between all these processes and lump them all together under some notion of "computation". But then, all mechanistic explanations would look very much alike, explaining every capacity in terms of inner "computations". Most of the explanatory power of mechanistic explanations, which depends on the differences between processes like digestion, refrigeration, and illumination, would be lost.

In other words, if mechanistic explanation is the same as computational explanation, then every artifact and biological organ is a computing mechanism. The brain is a computing mechanism in the same sense in which a stomach, a freezer, and a light bulb are computing mechanisms. This conclusion is counterintuitive and flies in the face of our scientific practices. It also trivializes CTC, which was designed to invoke a special activity (computation) to explain some special capacities (cognitive ones) as opposed to others. To avoid this consequence, we should conclude that

computation may well be a process that deserves to be explained mechanistically, but it should not be identified with every process of every mechanism. Computation is one special kind of mechanistic process among others.

As to functional analysis, to think of it as an explanatory strategy distinct from and autonomous from mechanistic explanation is a mistake. There is no such thing as an explanation of a phenomenon purely in terms of capacities or operations, without those capacities or operations being performed by some components of the system. (By the same token, there is no such thing as an explanation of a phenomenon purely in terms of components, without those components possessing capacities and performing operations.) A mechanistic explanation includes both structures (components) and functions (capacities, operations). In so far as anyone offers a "purely" functional analysis of a phenomenon, in which information about components is omitted, she is offering a *sketch* of a mechanism, which is waiting to be completed by adding the information about components. Once the information about components is added, the mechanism sketch turns into a more complete mechanistic explanation (Piccinini and Craver 2011).

This subsumption of functional analysis within mechanistic explanation is sometimes misunderstood as a call for maximal specificity—maximal amount of detail—in explanation. On the contrary, explaining requires providing all and only the information that is causally relevant to a phenomenon at a chosen level of generality, which requires abstracting away from irrelevant details of specific mechanisms that produce the phenomenon (Barberis 2013; Chirimuuta 2014; Levy and Bechtel 2013; Weiskopf 2011). Explanation also involves reconciling evidence produced using different operational definitions and different experimental protocols (Sullivan 2009). But the result of choosing an appropriate level of generality and reconciling evidence from different paradigms is *not* a *non*-mechanistic explanation. The result is a mechanistic explanation at the relevant level of generality, which may abstract away from and generalize over a wide variety of lower level mechanisms for the same phenomenon (cf. Boone and Piccinini 2015; Piccinini and Maley 2014).

13.6 Computing Mechanisms

In order to articulate an adequate notion of computational explanation we can make progress, as we did before, by looking at the explanatory strategies employed by the relevant community of scientists—specifically, computer scientists and engineers. In understanding and explaining the capacities of calculators and computers, computer scientists employ full-blown mechanistic explanation. They analyze computing systems into processors, memory units, etc., and they explain the computations performed by the systems in terms of the (teleological) functions performed by appropriately organized components. But computer scientists employ mechanistic explanations that are *specific* to their field: the components, functions, and organizations employed in computer science are of a distinct kind. If this is correct, then the appropriate answer to our initial question about the nature of

computational explanation is that it is a distinct kind of mechanistic explanation. Which kind?

There are many kinds of computing mechanisms, each of which comes with its specific components, functions, and functional organization. If we want to characterize computational explanations in a general way, we ought to identify features that are common to the mechanistic explanation of all computing mechanisms.

The modern, mathematical notion of (digital) computation, which goes back to work by Alan Turing ([1936] 1965) and other logicians, can be formulated in terms of strings of digits. For example, letters of the English alphabet are digits, and concatenations of letters (i.e., words) are strings of digits. More generally, digits are states that belong to finitely many types, which are unambiguously distinguishable by the mechanisms that manipulate them; strings are concatenations of digits. As I have argued elsewhere, *concrete digital computing systems* are *mechanisms whose (teleological) function is manipulating strings of digits in accordance with rules that are general*—namely, they apply to all strings from the relevant alphabet—*and that depend on the input strings (and perhaps internal states) for their application*.[4] A digital computational explanation, then, is a mechanistic explanation in which the inputs, outputs, and perhaps internal states of the system are strings of digits, and the processing of the strings can be accurately captured by appropriate rules.

This account applies to both ordinary computers as well as those classes of neural networks whose input–output functions can be analyzed within the language of computability theory (e.g., McCulloch and Pitts 1943; Minsky and Papert 1969; Hopfield 1982; Rumelhart and McClelland 1986; Siegelmann 1999). In other words, any connectionist network whose inputs and outputs can be characterized as strings of digits, and whose input–output function can be characterized by a fixed rule defined over the inputs (perhaps after a period of training), counts as a digital computing mechanism in the present sense.

But this analysis does not apply to so-called analog computers and other types of neural networks, which do not manipulate strings of digits. In order to cover such systems, we need a more general notion of computation than that of digital computation. Non-digital computation is *like* digital computation in that it is based solely on differences between different portions of the vehicles it operates on, without reference to any more concrete properties of the vehicles. In other words, all computation—whether digital or non-digital—is *medium-independent*.[5] But non-

[4]To be a bit more precise, for each digital computing system, there is a finite alphabet out of which strings of digits can be formed and a fixed rule that specifies, for any input string on that alphabet (and for any internal state, if relevant), whether there is an output string defined for that input (internal state), and which output string that is. If the rule defines no output for some inputs (internal states), the mechanism should produce no output for those inputs (internal states). For more details, see Piccinini 2015.

[5]Medium-independence entails multiple realizability but not vice versa. Any medium-independent vehicle or process is realizable by different media, thus it is multiply realizable. But the converse does not hold. Functionally defined kinds, such as *mousetrap* and *corkscrew*, are typically multiply realizable—that is, they can be realized by different kinds of mechanisms (Piccinini and Maley

digital computation is *different* from digital computation in that its vehicles are different from strings of digits. Therefore, we need to define computation in the generic sense in terms of vehicles that are more general than strings of digits.

A computation in the generic sense, then, is a process defined by a general rule for manipulating some kind of vehicle based on differences between different portions of the vehicle along some dimension of variation. A *computing system in the generic sense* is a *mechanism whose (teleological) function is manipulating some type of vehicle* (digital, analog, or what have you) *in accordance with a rule that is general*—namely, it applies to all vehicles of the relevant kind—*and that depends on the input vehicles (and perhaps internal states) for its application* (Piccinini 2015). A computational explanation in the generic sense, then, is a mechanistic explanation in which the inputs, outputs, and perhaps internal states of the system are medium-independent vehicles, and the processing of the vehicles can be accurately captured by appropriate rules.

With the present account of computation in hand, we are ready to discuss whether and how computation is related to cognition. Since we are interested in explaining cognitive capacities, we will focus primarily not on genetic or molecular neuroscience but on the levels that are most relevant to explaining cognitive capacities, that is, cellular and systems neuroscience.

13.7 Neural Mechanisms

Analogously to mechanistic explanation in other fields, mechanistic explanation in neuroscience is about how different neural components and their capacities are organized together so as to exhibit the activities of the whole. But mechanistic explanation in neuroscience is also different from mechanistic explanation in most other domains, due to the peculiar functions performed by nervous systems.

The functions of the brain (and more generally of the nervous system) may be approximately described as the feedback control of the organism and its parts.[6] In other words, the brain is in charge of bringing about a wide range of activities performed by the organism in a way that is sensitive to the state of both the organism and its environment. The activities in question include, or course, the cognitive functions as well as any functions involving the interaction between the whole organism and the environment, as in walking, feeding, or sleeping, and in addition a wider range of functions ranging from breathing to digesting to releasing hormones into the bloodstream.

2014). But most functionally defined kinds, including *mousetrap* and *corkscrew*, are *not* medium-independent—they are defined in terms of specific physical effects, such as catching mice or lifting corks out of bottles.

[6]The exact level of sophistication of this feedback control is irrelevant here. Cf. Grush (2003) for some options.

Different functions come with different mechanistic explanations, and feedback control is no exception. In order to be properly sensitive to the state of both the organism and its environment, the nervous system must collect and carry information (Adrian 1928; Garson 2003). In the present sense, carrying information means possessing internal variables (vehicles) that *reliably correlate* with what they carry information about (Piccinini and Scarantino 2011).

Given that nervous systems have this peculiar capacity of performing feedback control, their mechanistic explanation requires the appeal to internal states that reliably correlate with the rest of the body, the environment, and one another in appropriate ways. In order for a nervous system to control an organism, the internal variables must drive the activities of the organism in appropriate ways through effectors (muscles and glands). In order for the control to be based on feedback, the brain's internal variables must carry information about bodily and environmental variables. When a system has this kind of information-carrying and -processing *function*, we say that the system possesses and processes internal *representations* (Dretske 1988).[7]

The vehicles employed by brains to represent the body and the environment are, of course, all-or-none events, known as action potentials or spikes, which are generated by neurons. Spikes are organized in sequences called *spike trains*, whose properties vary from neuron to neuron and from condition to condition. Spike trains from one neuron are often insufficient to produce a functionally relevant effect. At least in large nervous systems, such as human nervous systems, in many cases spike trains from populations of several dozens of neurons are thought to be the minimal processing units (Shadlen and Newsome 1998). Mechanistic explanation in neuroscience, at or above the levels that interest us here, consists in specifying how appropriately organized trains of spikes from different neuronal assemblies constitute the capacities of neural mechanisms, and how appropriately organized capacities of neural mechanisms constitute the brain's capacities—including its cognitive capacities.

13.8 Neural Computation

We are now in a position to evaluate whether the nervous system performs computations and what kind of computation it performs. We have seen that within current neuroscience, the variables that are employed to explain cognitive capacities are spike trains generated by neuronal populations. The question of whether cognition has a computational explanation becomes the question of whether the neural processes that explain cognitions are computations. A positive answer presupposes that spike trains are computational vehicles.

[7]Ramsey (2007) criticizes Dretske's account of representation; Morgan (2014) defends its adequacy.

If we reformulate the original CTC using modern terminology, CTC was initially proposed as the hypothesis that spike trains are strings of digits and neural processes are digital computations (McCulloch and Pitts 1943; Piccinini 2004a). Furthermore, the neural computation literature makes references to literature and results from computability theory (which pertains to digital computation) as if they were relevant to neural computation (e.g., Churchland and Sejnowski 1992; Koch 1999). We can now evaluate the hypotheses that neural processes are computations and the stronger hypothesis that neural computations are digital by looking at what neuroscientists have empirically discovered by studying spike trains.

There are a couple of powerful reasons to conclude that neural processes are computations in the generic sense (Piccinini and Bahar 2013). First, neural vehicles—primarily, spike trains—appear to be medium-independent. That is to say, the functionally relevant properties of spike trains are spike frequency and spike timing, both of which are defined independently of the specific physical properties of spikes. Furthermore, nervous systems process spike trains in accordance with rules that are defined over the vehicles and responsive to the inputs and internal states of the system. Therefore, neural processes operate on medium-independent vehicles in accordance with appropriate rules, which is what we defined computation in a generic sense to be (Sect. 13.6).

A second reason for the same conclusion is that neural systems use spike trains to encode information and they process information by processing spike trains. Information is a medium-independent notion. In the present sense, carrying information about a variable means correlating reliably with that variable. Processing information in this sense means processing variables based on the information they carry, regardless of how such variables are physically implemented. Another route to this conclusion goes through the fact that neural vehicles must carry information about many physically different variables, such as light waves, sound waves, muscle contractions, and many more. The most efficient way to carry information about all such physically different sources and process such information in the service of controlling the organism is to establish an internal common currency—the relevant neural variables, the spike trains—that is defined independently of the specific physical properties of what they carry information about. That is what nervous systems do. Therefore, processing information in this sense means manipulating medium-independent variables in a medium-independent way. Again, that means carrying out computations in the generic sense.

So neural processes are computations in the generic sense. What about the more specific hypothesis that neural computations are digital? A close look at current neuroscience reveals that within biologically plausible models, spike trains are far from being described as strings of digits. There appear to be principled reasons why this is so.

Since strings are made of atomic digits, if neural computations were digital it must be possible to decompose spike trains into atomic digits, namely events that have unambiguous functional significance within the mechanism during a functionally relevant time interval. But the current mathematical theories of spike generation (Dayan and Abbott 2001; Ermentrout and Terman 2010) leave no room

for doing this. The best candidates for atomic digits, namely the presence and absence of a spike, have no determinate functional significance on their own; they only acquire functional significance within a spike train by contributing to an average firing rate. Moreover, spikes and their absence, unlike atomic digits, are not events that occur within well-defined time intervals of functionally significant duration.

Even if the presence or absence of individual spikes (or any other aspect of spike trains) could be usefully treated as an atomic digit, however, it would be unclear how they could be concatenated into strings, which is a necessary condition for them to be vehicles for digital computation. In order for digits to be concatenated into strings, it must be possible to determine unambiguously, at the very least, which digits belong to a string and which do not. But again, within our current mathematical theories of spike trains, there is no non-arbitrary way to assign individual spikes (or groups thereof) to one string rather than another.

Based on current evidence, then, neural computation is not digital. It is not analog computation either, because spike trains are not continuous signals like those used by analog computers but are made of discrete spikes. More precisely, what is functionally significant about spike trains is not the precise voltage values or the exact evolution of the voltage values during a time interval, but the spike frequency (and possibly spike timing) during that same interval. Therefore, neural computation is its own kind of computation—distinct from both digital and analog computation (Piccinini and Bahar 2013).

13.9 Explaining Cognition Computationally

I have argued that once the notion of computation that is relevant to CTC is in place, we should conclude that neural processes are sui generis computations. As far as we can now tell, based on our best neuroscience, cognitive capacities are explained by the processing of spike trains by neuronal populations, and the processing of spike trains is a kind of computation that is interestingly different from digital computation as well as analog computation.

Some philosophers may be tempted to reply that their favorite version of CTC is not threatened by neuroscience, because CTC is "autonomous" from neuroscience (e.g., Fodor 1997). According to this line, CTC is a theory at a higher, more "abstract" level than the neural level(s); it is a psychological theory, not a neuroscientific one; it is not directly constrained by the properties of the neural realizers. The autonomy reply goes against the spirit in which CTC was proposed, for CTC was originally proposed as a theory of how cognitive capacities are explained by *neural* mechanisms (McCulloch and Pitts 1943; Wiener 1948; von Neumann 1958). Anyone who takes this autonomy defense is liable to be criticized for expounding a reactionary theory—reactionary because immune to empirical revision (cf. Churchland 1981).

Most importantly, the autonomy defense is based on a faulty account of the relationship between psychology and neuroscience—the account that I rejected in Sect. 13.5. Any supposedly more "abstract" or "purely functional" explanation of cognition is, on close examination, a sketch of a mechanism, to be filled in with information about the components that perform that relevant activities or functions. Needless to say, the components are neural components, and the relevant evidence comes from neuroscience.

References

Adrian, E. D. (1928). *The basis of sensation: The action of the sense organs*. New York: Norton.
Barberis, S. D. (2013). Functional analyses, mechanistic explanations, and explanatory tradeoffs. *Journal of Cognitive Science, 14*(3), 229–251.
Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.
Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as scientific research strategies*. Princeton: Princeton University Press.
Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology* (6th ed., pp. 261–325). Minneapolis: University of Minnesota Press.
Boone, T., & Piccinini, G. (2015). "The cognitive neuroscience revolution". Synthese. doi:10.1007/s11229-015-0783-4
Bringsjord, S. (1995). Computation, among other things, is beneath us. *Minds and Machines, 4*, 469–488.
Burge, T. (1986). Individualism and psychology. *Philosophical Review, 95*, 3–45.
Chalmers, D. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science, 12*(4), 323–357.
Chirimuuta, M. (2014). Mazviita Chirimuuta, minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*.
Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.
Copeland, B. J. (2000). Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind-machine issue. *The Journal of Philosophy, XCVI*(1), 5–32.
Craver, C. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Oxford: Oxford University Press.
Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, MA: MIT Press.
Cummins, R. (2000). "How does it work?" vs. "What are the laws?" Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition*. Cambridge, MA: MIT Press.
Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.
Dennett, D. C. (1978). *Brainstorms*. Cambridge, MA: MIT Press.
Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.
Dreyfus, H. L. (1998). Response to my critics. In T. W. Bynum & J. H. Moor (Eds.), *The digital phoenix: How computers are changing philosophy* (pp. 193–212). Malden: Oxford, Blackwell.
Edelman, G. M. (1992). *Bright air, brilliant fire: On the matter of the mind*. New York: Basic Books.
Egan, F. (1995). Computation and content. *Philosophical Review, 104*, 181–203.

Ermentrout, G. B., & Terman, D. H. (2010). *Mathematical foundations of neuroscience*. New York: Springer.
Erneling, C. E., & Johnson, D. M. (2005). *The mind as a scientific object: Between brain and culture*. Oxford: Oxford University Press.
Fetzer, J. H. (2001). *Computers and cognition: Why minds are not machines*. Dordrecht: Kluwer.
Fodor, J. A. (1968). *Psychological explanation*. New York: Random House.
Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives, 11,* 149–163.
Fodor, J. A. (1998). *Concepts*. Oxford: Clarendon Press.
Fresco, N. (2014). *Physical computation and cognitive science*. New York: Springer.
Gallistel, C. R., & King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Malden: Wiley-Blackwell.
Garson, J. (2003). The introduction of information into neurobiology. *Philosophy of Science, 70*, 926–936.
Gerard, R. W. (1951). Some of the problems concerning digital notions in the central nervous system. In H. v. Foerster, M. Mead, & H. L. Teuber (Eds.), *Cybernetics: Circular causal and feedback mechanisms in biological and social systems. Transactions of the seventh conference* (pp. 11–57). New York: Macy Foundation.
Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
Glennan, S. S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 64*, 605.
Globus, G. G. (1992). Towards a noncomputational cognitive neuroscience. *Journal of Cognitive Neuroscience, 4*(4), 299–310.
Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy, 18*, 53–93.
Harnad, S. (1996). Computation is just interpretable symbol manipulation; cognition isn't. *Minds and Machines, 4*, 379–390.
Haugeland, J. (1997). What is mind design? In J. Haugeland (Ed.), *Mind design II* (pp. 1–28). Cambridge, MA: MIT Press.
Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*, 2554–2558.
Horst, S. W. (1996). *Symbols, computation, and intentionality: A critique of the computational theory of mind*. Berkeley: University of California Press.
Johnson, D. M., & Erneling, C. E. (Eds.). (1997). *The future of the cognitive revolution*. New York: Oxford University Press.
Koch, C. (1999). *Biophysics of computation: Information processing in single neurons*. New York: Oxford University Press.
Lucas, J. R. (1996). Minds, machines, and Gödel: A retrospect. In P. J. R. Millikan & A. Clark (Eds.), *Machines and thought: The legacy of Alan Turing*. Oxford: Clarendon.
Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science, 80*(2), 241–261.
Machamer, P. K., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.
Maley, C., & Piccinini, G. (forthcoming). The ontology of functional mechanisms. In D. Kaplan (Ed*.), Integrating psychology and neuroscience: Prospects and problems*. Oxford: Oxford University Press.
Marr, D. (1982). *Vision*. New York: Freeman.
Maudlin, T. (1989). Computation and consciousness. *Journal of Philosophy, 86*(8), 407–432.
McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 7*, 115–133.
Mellor, D. H. (1989). How much of the mind is a computer? In P. Slezak & W. R. Albury (Eds.), *Computers, brains and minds* (pp. 47–69). Dordrecht: Kluwer.
Milkowski, M. (2013). *Explaining the computational, mind*. Cambridge, MA: MIT Press.
Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
Morgan, A. (2014). Representations gone mental. *Synthese, 191*(2), 213–244.

Newell, A., & Simon, H. A. (1976). Computer science as an empirical enquiry: Symbols and search. *Communications of the ACM, 19*, 113–126.
Ó Nualláin, S., & Mc Kevitt, P. (Eds.). (1997). *Two sciences of mind: Readings in cognitive science and consciousness*. Philadelphia: John Benjamins.
O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
Penrose, R. (1994). *Shadows of the mind*. Oxford: Oxford University Press.
Perkel, D. H. (1990). Computational neuroscience: Scope and structure. In E. L. Schwartz (Ed.), *Computational neuroscience* (pp. 38–45). Cambridge, MA: MIT Press.
Piccinini, G. (2004a). The first computational theory of mind and brain: A close look at McCulloch and Pitts's 'logical calculus of ideas immanent in nervous activity'. *Synthese, 141*(2), 175–215.
Piccinini, G. (2004b). Functionalism, computationalism, and mental states. *Studies in the History and Philosophy of Science, 35*(4), 811–833.
Piccinini, G. (2007). Computational explanation and mechanistic explanation of mind. In M. De Caro, F. Ferretti, & M. Marraffa (Eds.), *Cartographies of the mind: Philosophy and psychology in intersection* (pp. 23–36). Dordrecht: Springer.
Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.
Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science, 34*, 453–488.
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283–311.
Piccinini, G., & Maley, C. (2014). The metaphysics of mind and the multiple sources of multiple realizability. In M. Sprevak & J. Kallestrup (Eds.), *New waves in the philosophy of mind* (125–152). Palgrave Macmillan.
Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics, 37*(1), 1–38.
Port, R. F., & van Gelder, T. (Eds.). (1995). *Mind and motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.
Putnam, H. (1967). *Psychological predicates* (Art, philosophy, and religion). Pittsburgh: University of Pittsburgh Press.
Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.
Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.
Ramsey, W. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.
Rubel, L. A. (1985). The brain as an analog computer. *Journal of Theoretical Neurobiology, 4*, 73–81.
Rumelhart, D. E., & McClelland, J. M. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.
Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge: Cambridge University Press.
Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.
Segal, G. (1991). Defence of a reasonable individualism. *Mind, 100*, 485–493.
Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *Journal of Neuroscience, 18*(10), 3870–3896.
Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*(438), 369–400.
Shagrir, O. (2006). Why we view the brain as a computer. *Synthese, 153*(3), 393–416.
Siegelmann, H. T. (1999). *Neural networks and analog computation: Beyond the Turing limit*. Boston: Birkhäuser.
Sullivan, J. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese, 167*, 511–539.
Taube, M. (1961). *Computers and common sense: The myth of thinking machines*. New York: Columbia University Press.

Thelen, E., & Smith, L. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.
Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. In M. Davis (Ed.), *The undecidable*. Hewlett: Raven.
van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy, XCII*(7), 345–381.
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
von Neumann, J. (1958). *The computer and the brain*. New Haven: Yale University Press.
Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese, 183*, 313–338.
Wiener, N. (1948). *Cybernetics or control and communication in the animal and the machine*. Cambridge, MA: MIT Press.
Wilson, R. A. (2004). *Boundaries of the mind: The individual in the fragile sciences*. Cambridge: Cambridge University Press.
Wright, C. (1995). Intuitionists are not (Turing) machines. *Philosophia Mathematica, 3*(3), 86–102.

Chapter 14
Representational Development Need Not Be Explicable-By-Content

Nicholas Shea

Abstract Fodor's radical concept nativism flowed from his view that hypothesis testing is the only route to concept acquisition. Many have successfully objected to the overly-narrow restriction to learning by hypothesis testing. Existing representations can be connected to a new representational vehicle so as to constitute a sustaining mechanism for the new representation, without the new representation thereby being constituted by or structured out of the old. This paper argues that there is also a deeper objection. Connectionism shows that a more fundamental assumption underpinning the debate can also be rejected: the assumption that the development of a new representation must be explained in content-involving terms if innateness is to be avoided.

Fodor has argued that connectionism offers no new resources to explain concept acquisition: unless it is merely an uninteresting claim about neural implementation, connectionism's defining commitment to distributed representations reduces to the claim that some representations are structured out of others (which is the old, problematic research programme). Examination of examples of representational development in connectionist networks shows, however, that some such models explain the development of new representational capacities in non-representational terms. They illustrate the possibility of representational development that is not explicable-by-content. Connectionist representations can be distributed in an important sense, which is incompatible with the assumption of explanation-by-content: they can be distributed over non-representational resources that account for their development. Rejecting the assumption of explanation-by-content thereby opens up a more radical way of rejecting Fodor's argument for radical concept nativism.

Keywords Conceptual development • Nativism • Connectionism • Parallel distributed processing • Representational development

N. Shea (✉)
King's College London, London, UK
e-mail: nicholas.shea@kcl.ac.uk

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_14

14.1 Introduction

How can new representations be acquired? When that question is asked about new concepts, Fodor famously argued that hypothesis testing is the only option (Fodor 1975, 1981). That led him to embrace radical concept nativism.[1] Many objectors have pointed out that hypothesis testing is not the only candidate learning mechanism (Carey 2009; Cowie 1999; Laurence and Margolis 2002), showing how existing representations can be involved in the acquisition of new ones without the new representations thereby being structured out of the old (Margolis 1998; Rupert 2001).[2] This paper argues that the point runs deeper. The development of a new representation need not be explicable in content-involving terms at all. It may proceed by putting together non-representational resources in such a way as to constitute an entirely new representation.

Most answers to Fodor's challenge rely on existing representational resources in explaining the development of a new representation type. For example, a new natural kind term can be acquired by recognising salient properties of an object, connecting them with a new internal symbol, and combining that with an essentialist principle (Margolis 1998). Such accounts are important in their own right, but do not challenge the tacit assumption that, if innateness is to be avoided, representational development must consist in a series of stages or transitions that are explicable in terms of the semantic content of the representations involved. Connectionist systems offer an excellent illustration of the more radical claim. Fittingly, Fodor's robust challenge to the usefulness of connectionist modelling brings out the central role that an assumption of explanation-by-content is playing in his arguments.

One of Fodor's challenges to connectionism is characteristically pithy: connectionists' defining characteristic is a commitment to distributed representations – over *what* are they distributed? At best, Fodor argues, connectionists can be saying nothing more than that connectionist representations are distributed over further representations, those found at the level of their processing units (Sect. 14.2). But if so, connectionism is nothing new. It is just a restatement of the old programme in which new representations are complex constructs out of innate primitives. For present purposes we can concede to Fodor that that programme has been unsuccessful, especially as an account of lexical concepts.

Fodor's challenge forces us to be very precise about what connectionism should say about the development of new representations. This paper examines some examples to show that even static feedforward PDP models provide a genuinely novel way of seeing the development of new representational resources. They show us how there can be a non-representational but informational explanation of the

[1] Fodor has since retreated somewhat from that position: Fodor (1998, 2008), including becoming more sympathetic to alternatives to hypothesis testing as an account of concept acquisition (Fodor 2008, pp. 162–168).

[2] Fodor now accepts that such processes do not depend on hypothesis testing (Fodor 2008), but still argues that they form part of a creature's innate conceptual endowment (2008, pp. 163–164).

development of entirely new primitive representations (Sect. 14.3). Connectionism's interesting distribution claim is that representations can be distributed over the entities that account for their development (Sect. 14.4). That is perhaps the most important theoretical insight offered by PDP modelling.

At first sight, Fodor's challenge to connectionism appears to be independent of his argument for radical concept nativism. On examination, they both turn out to depend upon the same tacit assumption: that if the development of a new representation is to fall within the ambit of psychology, it must occur as a rational transition or inference between semantic items, explicable in virtue of their contents (it must be 'explicable-by-content'). Connectionist models of representational development show that there can be a psychological explanation of the transition to a new representation that is not an explanation-by-content. That point generalises into a tactic for answering Fodor's puzzle of radical concept nativism in other cases (Sect. 14.5). However, the nativists had a good point too, because it would be a mistake to think, as many connectionists have, that the acquisition of new representational resources is to be explained in terms of their contents. The grain of truth in Fodor's radical concept nativism is that, for very many of the representations that are responsible for intelligent behaviour, their development is not explicable-by-content. His mistake was to conclude that they must therefore be innate.

Fodor's innateness argument concerns concepts. Concepts are one species of mental representation. They are constituents of complete thoughts. Complete thoughts have associated conditions of satisfaction or truth/correctness. Concepts, taken alone, do not. Fodor hypothesises a language of thought, such that all mental representations with conditions of truth/correctness or satisfaction are formed out of constituent concepts (but cf. Fodor 2007). Representationalism is more permissive allowing, for example, that there are representations which have satisfaction conditions that we would express using a sentence (eg, *there is a snake on the ground, climb a tree*), but where the representation itself contains no constituent structure (nothing in the representation corresponds separately to *the ground, snakes* or *climbing*). I will use the term 'non-conceptual representation' for psychological states with correctness or satisfaction conditions but no constituent structure. Non-conceptual representations are probably needed to understand many classes of PDP model, as well as many psychological phenomena. Although Fodor talks about concepts, the considerations he canvasses in support of nativism are equally applicable to non-conceptual representations. So I will move freely between talking about concepts and representations in general.

14.2 Fodor's Argument Against Connectionism

Fodor has several objections to connectionism. The most well-known is that connectionist models cannot explain the systematicity and productivity of thought (Fodor and McLauglin 1990). Also prominent is the claim that connectionists cannot avoid an unacceptably holist theory of the content of distributed representations

(Fodor and Lepore 1992). My focus is a third objection: that connectionism is nothing new. According to Fodor, all it has to offer is the standard idea that some mental representations are structured out of others, coupled with an outdated associationism about mental processing (Fodor and Pylyshyn 1988). Supposed theoretical insights, like the idea of distributed representation, and of learning by modulation of connection strengths, are simply new pieces of terminology for old ideas, terminology that obscures the failings inherent in treating lexical concepts as structured, but does nothing to address them.

The argument can be formulated as a dilemma.[3] Connectionists claim that mental representations are distributed. What type of objects are they distributed over? Characteristically, Fodor offers a dichotomy:

Version One: Mental representations are distributed over neurons.[4]
Version Two: Some mental representations are distributed over others.[5]

According to Version One, connectionism is just a claim about how mental representations are realised. Any psychological theory must be realised somehow in physical brains. No one thinks that all mental representations correspond to individual neurons or "grandmother cells". To be of theoretical interest, connectionists need to make some claims that connect with the explanatory level of psychology. So Fodor argues.

PDP modellers themselves are unlikely to accept the Version One characterisation. Of course when the models are applied to the real world, representations will be realized in multiple neurons. But the units over which representations are distributed in the models are not neurons. PDP modellers often explicitly eschew a commitment to a 1–1 correspondence between processing units and neurons. Version One therefore does not capture the force of PDP's distribution claim.

Fodor offers Version Two as the only alternative. But this is just a familiar story about structured representations. Connectionists' distributed representations are merely some kind of complex constructs out of the representations that are their constituent units. For example, Fodor and colleagues interpret Churchland's "state space semantics" (Churchland 1998, 2012) as treating individual hidden layer units as representing complex microfeatures, with a distributed pattern of activation having its content as some kind of complex weighted conjunction of those microfeatures (Fodor and Pylyshyn 1988, pp. 19–21; Fodor and Lepore 1999, p. 391). Distributed representations are structured out of the representations over which they are distributed, and the content of a distributed representation is fixed by the contents of the constituent units.

[3]This objection to both the versions of connectionism offered here has been raised by Fodor in many places, e.g. in Fodor and Pylyshyn (1988) and Fodor and McLaughlin (1990). The formulation explicitly in terms of a dilemma is found in Fodor (2004), a draft paper posted on the New York University website.

[4]Fodor and Pylyshyn (1988), pp. 19 & 64–68.

[5]Fodor and Pylyshyn (1988), pp. 19–21.

Fodor rejects Version Two on the basis that constructing concepts out of pre-existing representations is a failed research programme. He argues that there are no plausible definitions for most lexical concepts, and that neither prototypes nor exemplars compose in the way that is required by a compositional semantics. Connectionists can, and often do, object at this stage. Fodor's objections to prototype and exemplar theories may be surmountable. Or the connectionist's way of constructing distributed representations out of the contents of individual nodes may be different in important respects, so as to overcome extant objections. Furthermore, "constructivist" neural networks side-step the worry about constructing distributed representations out of existing microfeatures since they allow for the recruitment of new hidden units that previously played no role in the network (Mareschal and Schultz 1996; Quartz and Sejnowski 1997).

These lines of reply to Fodor are familiar. They may be the best way to characterise some classes of connectionist models. But there is another answer available too. Fodor's dilemma presupposes that there are only two candidates for the entities over which mental representations are distributed: neurons or further representations. To have any bite connectionists do indeed have to tell us what it is that distributed representations are distributed over. But Fodor has offered connectionists a false dichotomy. To see that there is another possibility we must first get on the table a positive account of the development of new representations in connectionist systems.

14.3 Developing New Connectionist Representations

Even static, programmer-designed neural networks can develop novel representations. This section gives an account of how. In particular, it shows how familiar training algorithms can transform a system without hidden layer representations into one that has new representational capacities.

The basic idea is that there are connectionist learning algorithms that transform (non-semantic) information into representation. Before a connectionist system has been trained, the units of its hidden layers, and perhaps its input layer too, can be merely information-carriers. Their tokening will correlate with various features of the items coded as input. When the instantiation of some property F by an object changes the probability of the instantiation of another property G by an object, we can say that F carries *correlational information* about G. Correlational information is ubiquitous. Representation is something more substantial. The fact that single units and distributed patterns of activation carry correlational information (about all sorts of affairs) does not imply that they have representational content. Typically, it is only after training that distributed patterns of activation have the right properties to have genuinely representational content (truth conditions, satisfaction conditions, etc.). Of course, there is no agreement as to exactly what more is needed, but all sides agree that bare correlational information is not sufficient for representation.

Connectionists need not think of individual units as being representational at all. Indeed, Chalmers (1992) takes that to be characteristic of connectionist models: the items over which computational processes are defined are more fine-grained than the lowest level at which representational contents are properly attributable to states of the system. (Some networks *are* designed to have individual units as representational, e.g. the semantic networks of Quillian 1967.) Connectionists have tended to accept that individual units represent something (e.g. complex microfeatures), when they need not. That takes connectionism towards the Version Two interpretation and its attendant problems. In fact, in many networks there is no reason to think of individual units as representational at all.

An example is the colour classification network of Laakso and Cottrell (2000). One way of coding the inputs there proceeded as follows. For a given colour patch, reflectance readings from a spectrophotometer were taken at 12 places on the electromagnetic spectrum (between wavelengths of 400 nm and 700 nm, at 25 nm intervals). The readings were normalised to the range 0–255 and converted into binary format (eg, 11010011), giving a list of 12 binary numbers for each colour sample. This list of binary numbers was converted into a 96-dimensional vector of 0s and 1s to act as input vector ($96 = 12$ binary numbers of 8 digits). In that coding it is very hard to see an individual one of the 96 input units as representing anything at all. The particular 0 or 1 it carries makes sense only as part of a binary representation of magnitude that is distributed across 8 units. So even at the input layer there are cases where the individual units are not representational and only distributed patterns of activation are.

The case is even clearer when we come to hidden layers. There are good reasons, in many classes of model, not to treat single units of a hidden layer as representational. It is a mistake to concede that individual hidden layer units represent some kind of complex microfeatures. Shea (2007b) describes a class of connectionist systems in which individual hidden layer units are not representational. The networks do feature distributed representations. However the representations are distributed over network units, not over further representations. In other cases, the representations may be dynamic attractors in activation space (Clark 2001, p. 135, e.g. McLeod et al. 2000), making the representational level even further removed from the individual units in a single layer. Importantly, an explanation of how new representations develop (in those cases, how clusters or dynamic attractors develop in hidden layer state space) *is* given at the level of individual units.

So individual units in a connectionist network may not be representations: individual hidden layer units are unlikely to have representational content before training, and in many cases individual input layer units do not have representational content at all. But notice that each individual unit will carry correlational information, both before and after training (indeed, units will carry information about very many properties of the samples that have been coded into inputs). In many

connectionist networks, training encourages the network to form representations.[6] Some simple examples bring out the point. Competitive networks use unsupervised learning to find clusters in the inputs on which they are trained (Rolls and Treves 1998, Ch. 4). Unsupervised learning in auto-associative networks can also serve to identify the central tendencies or prototypes found in a range in input data, even where the prototype itself was never encountered in training (Plunkett and Sinha 1992). Both start with bare correlational information and end up with vehicles (clusters, prototypes) that are plausibly representations.

The point about the development of new representations can be made most starkly in networks in which there is no representation at all at the level of individual units before training, like the examples above. But that is not essential. The absence of initial representations just serves to make it obvious that the way that new representational capacities develop is not explicable-by-content. Representational development in these cases is a matter of using statistical learning to build mere information-bearers into representations. In other cases, pre-existing resources that are representations play a role in this process. What is crucial is that their role is merely causal. The way a new representational type develops, at a hidden layer say, depends on the correlational information carried by input units and hidden units, but there is no rational or content-based explanation of the transition from initial resources to new representations. Existing resources like the input units are relied on for the correlational information they carry, on which the connectionist training algorithm can act; but the story of the building of the new representational capacities is causal-correlational, not representational.

14.3.1 Application to an Example

To discuss a widely-known example, Sejnowski and Rosenberg's (1987) NETtalk network was trained using supervised learning to map English text to phonetic representations of its pronunciation. Where networks undergo supervised learning, clusters may form in hidden layer state space, leading to new distributed representations at the hidden layer. In NETtalk, before training there were no relevant partitions or clusters in hidden layer state space (although distributed patterns of activation would necessarily have carried some correlational information from the outset). The result of training the network to produce correct representations of phonemes at the output layer was that the network learnt to categorise inputs into vowels and consonants at the hidden layer on the way.

[6]There are many examples in which learning in connectionist systems creates attractors or clusters in state space (Churchland and Sejnowski 1992; Rupert 1998, 2001; Tiffany 1999). If there are reasons to see those attractors as being representations, then this is a process of turning information into representation.

According to two representative theories of content, asymmetric dependence theory and infotel semantics, this process leads to the creation of new representations out of non-representational resources. Learning a new representation of Cs is a matter of acquiring a new mental item R with the right properties firstly, to count as a mental symbol, and secondly to have the content C. According to Fodor's asymmetric dependence theory of content, having a representation R with content C is a matter of having a mental symbol whose tokening covaries with the presence of Cs, and of asymmetric dependence: to the extent that the tokening of R also covaries with any other property C*, it would not so-covary if R did not also covary with C (Fodor 1990). Call the mechanism which puts R in the right relation of causal covariation and counterfactual dependence with C a *sustaining mechanism* (Cowie 1999, p. 101; Laurence and Margolis 2002). Fodor's theory of content has faced many objections, and Fodor doesn't seem particularly keen on it himself,[7] but taking it at face value, learning a new representation of C is just a matter of going through a psychological acquisition process which results in a sustaining mechanism that connects a new symbol type R with Cs (with the appropriate causal profile).

Applying the asymmetric dependence theory to NETtalk, it is reasonably clear that there is no representation in the hidden layer at the outset, when connection weights are set randomly or arbitrarily. From the outset, both input and hidden layer units will carry a variety of correlational information, but there is no basis for thinking that there are any relations of asymmetric dependence amongst these correlations. After training, activation of the vowel partition of hidden layer state space correlates with presentation of a vowel to the network. It also correlates with other properties of the stimulus, say with the stimulus being a letter with a certain disjunctively-specified shape S. But that correlation is plausibly asymmetrically dependent on the correlation with vowels – were it not for the correlation with vowels, which is a useful intermediate to the classification made at the output layer, the network would not have arrived at a correlation with shape S.

For contrast, we can also assess the representational contents in NETtalk using infotel semantics (Shea 2007a), a modification of teleosemantics (Millikan 1984; Papineau 1987). Infotel semantics looks at the way a representation is used, as well as the way it is produced, in fixing its content. Out of all the correlational information carried by a putative representation, it focuses on the correlation that accounts for the system's having been trained (or evolved) to behave as it does (as argued by Dretske 1988; Ryder 2004 deploys a related idea). Applied to PDP models, this will deliver as content a condition specific to each representation-type, such that keeping track of that condition is what enables the network to produce correct outputs (where correctness is the standard against which the learning algorithm was trained).

[7] 'I assume that intentional content reduces (in some way or other, but, please, don't ask me how) to information; this is, I suppose, the most deniable thesis of my bundle.' (Fodor 1994, p. 4). 'If you want an externalist metaphysics of the content of innate concepts that's not just bona fide but true, I'm afraid there isn't one "yet".' Fodor (2001), p. 137.

Applied to NETtalk, infotel semantics implies that the output layer represents phonemes: in the course of training, the modeller took the units to represent phonemes, using that as the standard against which to generate an error signal. At the hidden layer, before training we have only correlational information. After training we have a partition of activation space into two groups of distributed patterns. Each correlates with a relevant feature of the input (vowel vs. consonant), and that distinction is consumed in downstream processing as a means to further phonetic categorisation. So according to infotel semantics there are representations in the hidden layer after training, but there is only correlational information before.

Notice that in this case, although there are representations at the input layer throughout (of strings of text), their role in fixing the content of the new representations (vowel vs. consonant) formed at the hidden layer is merely causal. It is not as if *vowel*, say, has been defined as some complex property of strings of text. Instead, input encodings of words into text strings serve as the causal basis for a sustaining mechanism that connects clusters at the hidden layer with properties of words. There is no explanation-by-content of the transition from a non-representational hidden layer, before training, to representations of vowels and consonants at the hidden layer, after training.

14.4 Connectionism's Interesting Distribution Claim

Armed with this account of the way connectionist networks can develop novel representations out of non-representational resources, we can return to Fodor's dilemma: *what* are connectionist representations distributed over?

In cases like the hidden layer representations in NETtalk, distributed representations are the lowest level of grain at which representational contents are properly attributable to the system. This fits Chalmers' (1992) observation that computational processes go on at a more fine-grained level (individual units) than the lowest level at which representations are found (distributed patterns of activation). Similarly, the story about how new representations develop is located at the more fine-grained level of single units. It is a recognisably psychological story, a form of statistical learning based on the way activation of units correlates with external features, and on correlations in activation between units.

We can clearly distinguish between the two levels of grain: one at which representations are found, another which figures in an account of the development of new representations.[8] That is, we can distinguish between ways of carving the network up into individuals for two different purposes:-

Obj1. Vehicles of representational content – individuals that figure in a representational explanation of the synchronic online operation of the trained system.

Obj2. Developmental units – individuals that figure in an explanation of the development of new representations.

[8]Tiffany (1999) and Shea (2007b) make parallel claims about the vehicles of content.

Fodor's argument against connectionism assumes that Obj1 = Obj2. But connectionists can make a much more interesting claim: that Obj1 are distributed over Obj2. The objects over which representations are distributed are not further representations (Version Two connectionism). Nor are they something merely implementational like neurons (Version One connectionism), since Obj2 are individuals which *do* figure in a psychological explanation (of the development of new representations). The interesting connectionist claim is that representations can be distributed over the resources that lead to their development. That is a clear sense in which connectionists' commitment to distributed representations *is* something new. It breaks away from an assumption that is deeply entrenched in classical computational models – that development of new representations must take place over existing representational resources. In this way, connectionist modelling has furnished cognitive science with a genuine insight, opening up a previously unexplored portion of logical space.

There would be good reason to assume Obj1 = Obj2 if we were committed to the idea that the development of a new representation must be explained in contentful terms: as an inference or rational transition that makes sense in the light of the semantic content of the objects involved in that transition. That is to reject the possibility that individual units may have a causal role in the development of a new distributed representation in virtue of the correlational information they carry, not constituting its content directly, but instead forming a sustaining mechanism which gives rise to its content. That is, Fodor's dichotomy implicitly assumes that the development of a new representation must be explicable-by-content:

Assumption of explanation-by-content

Whenever it occurs by a psychological process, the development of a new representation must consist of a transition from existing representational resources to the new representation, explicable in terms of the contents of the respective representations.

When new representations develop in PDP models in the way analysed in the previous section, it is clear that Obj1 are distributed over Obj2. The identification of Obj1 with Obj2 is a substantive assumption that has been implicitly constraining theorising in cognitive science. It is motivated by the assumption of explanation-by-content. That assumption also underpins Fodor's strong innateness claims, as we shall see in the next section.

14.5 Avoiding Fodor's Argument for Radical Concept Nativism

But what about Fodor's argument for radical concept nativism? We have seen how even static connectionist models can account for the development of entirely novel representations. These novel representations are not innate: PDP models offer an

account of their development, and do so in recognisably psychological terms.[9] How, then, is Fodor's argument avoided?

At first pass, Fodor's argument that connectionism offers nothing new seems quite separate from his argument for radical concept nativism. In this section we will see that Fodor's argument for radical concept nativism is in fact underpinned by the same assumption that lay behind his identification of Obj1 with Obj2 in the last section. Connectionism's insight is to show why that assumption can be rejected. In this section we spell out how doing so side-steps Fodor's nativism puzzle.

Fodor argues that concepts are either constructed from primitives or they are innate. His view is that most lexical concepts – concepts at the level of grain of individual words – are not constructed out of primitives. So they are innate, which is to say that they are not acquired via a learning process. Why does Fodor think that learning can only consist in constructing new representations out of existing ones? Not all ways of acquiring a new representation count as learning. Neither a bump on the head nor clever neurosurgery are learning processes, so if new representations could be acquired in either of those ways they would not be learnt. By contrast, setting a parameter for a grammatical principle, detecting a correlation, and constructing a new prototype based on experience are all clear cases of learning. Fodor argues that they all involve testing a hypothesis about what is the case: that the ambient grammar is head-first, that A correlates with B, that birds typically have feathers. To test a hypothesis against experience, the learner has first to be able to represent the hypothesis. So hypothesis testing cannot be a way of acquiring genuinely new representational resources, ones whose expressive power extends beyond contents that can be constructing out of pre-existing representations.

The standard response is that not all learning mechanisms are forms of hypothesis testing. That answer is correct, but it is incomplete, because it doesn't tell us what learning processes look like that are not hypothesis testing.[10] Fodor's move equating learning with hypothesis testing is not just an observation about what learning happens to consist in. It runs deeper. The claim is that learning can only consist in rational transitions between representations (Fodor 1975, p. 36). If that were right, then a person would indeed need to be able to formulate a claim before they could learn that it was true, which would exclude a learning-based account of the acquisition of entirely novel representations. That presents a puzzle, since it is implausible that my concepts of a carburettor (CARBURETTOR) or of my friend John (JOHN) are innate.[11]

[9]The concept of innateness is notoriously problematic (Mameli 2008). Fodor's central concern is whether concepts are learnt (Fodor 1975, 1991, 1998, 2008; Cowie 1999; Samuels 2002), so here I will take it that innate representations are not learnt or otherwise acquired by a psychological process and that they admit of a poverty of the stimulus argument (Shea 2012a, b).

[10]Margolis (1998), Rupert (2001), Laurence and Margolis (2002) and Carey (2009) give detailed accounts of forms of concept learning that are not a matter of hypothesis testing; as has Strevens (2012) since the present paper was written.

[11]Fodor has softened slightly in more recent work. First he allowed that concepts themselves may not be innate – what is innate is, for each concept, a domain-specific disposition, specific to each

Several authors have suggested a strategy for answering Fodor's innateness puzzle (Macnamara 1986; Margolis 1998; Rupert 2001; Laurence and Margolis 2002). Assume that what makes a representation have the content it does is wholly or partly determined by its causal relations with things in the world. Such sustaining mechanisms for a representation R may depend, causally, on other representations R* without the content of R being determined by the content of the R* – R's content is fixed more directly by its causal relations with things in the world. R can then be atomic, neither structured nor constructed out of the R*. The anti-nativist tactic is to give a psychological story in which existing representations R* come to form the sustaining mechanism for a new representation R, where the process of forming the new representation type R is described merely causally, not as a content-driven process like inference.

The key to this strategy is that not all learning consists in rational transitions between representations. Fodor's commitment to explanation-by-content closes off that option (driving him toward innateness). And we can see why Fodor would think learning is restricted in that way. The central insight of cognitive science is the viability of content-based explanation – the explanation of behaviour in terms of rational transitions between mental representations. The reality of these mental processes is vindicated by causal transitions between representation tokens in virtue of their form, but explanatory purchase is achieved by describing such representations in terms of their content. Rational transitions between contentful representations are the very core of the representational theory of mind (and its offshoot, the computational theory of mind / language of thought). So it is natural that Fodor should think that all psychological processes must consist in transformations between mental representations that are explicable in terms of the content of those representations.

If all learning processes were like that, then Fodor would be right to claim that any way of acquiring new representations that did not relate them to existing representations would necessarily lie outside the explanatory ambit of psychology. We have seen a first response to Fodor in accounts where the development of new representations depends upon existing representations without the new representation being structured or constructed out of the old (Margolis 1998; Laurence and Margolis 2002; Rupert 2001). But we can go further and reject the deeper underpinnings of Fodor's argument if we can reject the assumption of explanation-by-content entirely.[12] We must show how there can be instances of learning that are susceptible to a recognisably psychological explanation, but which do not fit within

such concept, to acquire that concept (Fodor 1998). But this still leaves Fodor postulating an innate domain-specific ability to develop DOORKNOB as a result of interaction with doorknobs. He has since added that the innate endowment might determine the geometry of neural attractor landscapes that realise concepts (Fodor 2008, p. 164). The worry remains that far too much is being taken to be innate. For simplicity, this paper considers only Fodor's earlier innateness claim.

[12]Fodor has more recently accepted that these accounts of concept learning do not involve hypothesis testing (Fodor 2008, pp. 163–167), and even that there is a 'jump' from the existing representations that are involved in creating the prototype: 'we jump, by some or other "automatic" process, from our stereotypes to our concepts' (2008, p. 164). However, he does not draw the moral

the standard mould where the outcome (a new representation) can be explained as a rational transition from existing representations. The transition to a genuinely novel contentful item cannot itself be susceptible to explanation in terms of content.

Our account of the development of new representations in the PDP models in Sect. 14.3 above is an existence proof that there can be such cases. It escapes Fodor's argument for radical representational nativism by rejecting his implicit commitment to explanation-by-content. That commitment can now be seen to lie behind both his radical concept nativism and his rejection of connectionism. But once PDP modelling has opened up this portion of logical space, it becomes clear that other cases of representational development should be understood in the same way. Shea (2011) has argued that Carey's influential account of children's development of the concept of natural number (Carey 2009) also involves a step that is not explicable-by-content.[13] Below I offer an example that goes beyond connectionism to illustrate that this could be a more general phenomenon.

14.5.1 Face Recognition

Morton and Johnson's (1991) theory of the development of face recognition furnishes a further useful example of how acquisition could fashion representations out of purely non-representational resources. Tested 30 min after being born, infants show a tendency preferentially to look at moving stimuli that have a configuration of two blobs over a third blob, something like this:

This tendency seems to be innate, in the sense that no learning is involved in the infant coming to have the looking bias. A poverty-of-the-stimulus argument can be made about it. The infant's disposition preferentially to track this category of inputs (perhaps driven by a subcortical visuomotor pathway) implicitly carries the information that such stimuli are worth attending to and learning about. If we ask where *that* information came from, we have to appeal to the infant's evolutionary history, not its individual experience. We can suppose that the bias is adaptive – it works well in the kinds of environments infants are likely to find themselves in. The adaptive match between behavioural bias and usual environment is due to evolution, not individual learning.

---

that there are psychological acquisition processes that are not explicable-by-content. He argues that the way this process works is due to innate constraints (2008, p. 164).

[13]Carey also observes that the child makes a 'leap' when drawing a parallel between the operation of adding one object in the object file system and the process of counting on to the next item in the (initially uninterpreted) sequence of counting words. Shea (2011) argued that this is the step at which Fodor's argument is circumvented, and that this step is not explicable as a rational transition from the content of pre-existing representational resources.

The infant's unlearnt behavioural bias is then sufficient to give a second system the input it needs to learn to reidentify individual faces. Through being given the right kind of input, this learning system has the chance to extract the statistical properties that distinguish one face from another and the statistical invariants that signify the same face again. Once trained up, the second system also implicitly encodes information: a rich store of information about which features indicate the same face (John, say). Unlike the information in the initial visual tracking tendency, this latter match between system and environment is not due to evolution, but to individual learning (from the experience of seeing John).

What contents are represented at these two stages of development? The answer depends upon the correct theory of content, about which there is no consensus, so I will again deploy asymmetric dependence theory and infotel semantics. At birth infants have the capacity to detect moving blobs and certain configurations of blobs. Some internal state driving their looking behaviour covaries roughly with the presence of faces, but there do not seem to be asymmetric dependencies between the various kinds of information carried or, if there are, it is the capacity to detect faces that looks to be asymmetrically dependent on the capacity to detect configurations of blobs, rather than the other way round. So, according to Fodor's theory of content, infants do not represent faces at the outset.[14] As a result of learning, the infant comes to be able to reidentify a particular individual, John say, by his face: the infant categorises together a variety of different views of John, and can engage in John-relevant behaviour as a result. So the result is some internal vehicle which correlates with John, and may well have the right asymmetric dependence properties to count as a representation of John. Thus, according to asymmetric dependence, the infant initially has no representation of faces at all, but then comes to have the ability to represent John by his face.

Infotel semantics also delivers the result that the capacity to represent John is not innate. Since the visuomotor tracking bias present at birth seems to have the function of enabling learning about faces, it plausibly carries the content *that's a face, look at it*, even though it is only able to identify faces very roughly at that stage. So infotel semantics suggests that this basic capacity to represent faces is innate. The capacity to represent the particular individual John is not innate. Although, even at birth, there are features of the visual signal that correlate with the presence of John, these are not deployed by consumer systems in a John-relevant way. Only once learning has taken place, so that the infant can reidentify John and thereby engage in John-relevant behaviour, will infotel semantics deliver any representations of John. Thus, according to both theories of content, the capacity to represent John is not present initially, but only arises after the second system has done its job.

We have offered a psychological account of the development of the ability to represent John, but the transition to having a representation of John is not explicable-by content, whether asymmetric dependence or infotel semantics is the right theory

[14] Since it is not clear how the relevant counterfactuals are to be assessed, it is hard to reach definitive conclusions about how the asymmetric dependence theory will apply to specific cases.

of content. The capacity to represent lines and blobs figures only causally in the development of the sustaining mechanism for the infant's later representation of John. Having played its developmental role in selecting appropriate input, the initial visual tracking tendency plays no causal role in the synchronic operation of the sustaining mechanism (Johnson et al. 1991).

The developmental transition is not explained-by-content. Instead, it makes use of resources characterised in terms of correlational information, which do not have the characteristics needed to count as representational, and builds them into sustaining mechanisms that do count as representational. Even before learning, the visual signal carries information, in the correlational sense, about particular faces. There is something about the visual signal which correlates with looking at John, say – that is why statistical learning about individual faces works. On no view does this correlational information, present in some complex form in the visual signal, count as representational at the initial stage. But these information-bearers play a causal role in the development of the mature ability to reidentify John. The story of that developmental transition is an account of information-bearers being built up into a sustaining mechanism for a new representation. It is a psychological story, but it is not explanation-by-content. It is a psychological account of the creation of content.

Why is this developmental transition an instance of learning, rather than mere triggering or maturation? Because it is a psychological process that involves extracting information from the environment. The infant comes to represent a particular individual, John, by interacting with John. If that were just triggering, it would be a mere accident that causal intercourse with John was needed to trigger maturation of the infant's concept of John.[15] By contrast, according to Morton and Johnson's theory there is a very obvious reason why the ability to recognise John depends upon seeing John: because the learning process works by picking up statistical properties in visual signals that come from John – properties that go with its being the same face again.

One way of confirming that the mature representation of John is not innate is by deploying a poverty of the stimulus argument. A poverty of the stimulus argument is available about the neonate's ability selectively to track things which tend to be faces. So that ability is plausibly innate. But there is no poverty of the stimulus argument available about the infant's later ability to recognise John. The infant's John-recognition device implicitly encodes a wealth of information about which properties distinguish John's face from other faces and which properties are invariant over different views of John's face. The infant does not rely on its evolutionary history for that information (and could not), but extracts that information from its experience of interacting with John. So if Morton and Johnson

[15]Fodor says that interaction with doorknobs is needed to trigger the DOORKNOB concept because being a doorknob is a response-dependent property Fodor (1998). Whether or not that response works for DOORKNOB, it is implausible that being John (a particular person) is a response-dependent property.

are right, the infant's capacity to represent individual faces is not innate. Whether or not they are right, their theory provides a detailed example of how genuinely novel representations could be learnt.

In the last section we saw that PDP modelling of representational development opens up a new portion of logical space for cognitive science to explore: that representations are distributed over the resources that account for their development, breaking the link between psychological explanation and explanation-by-content. In this section we saw that the same tactic gives us a more general answer to Fodor's puzzle about representational innateness.

14.6 Conclusion

Fodor assumes that all psychological processes, including concept acquisition, are intentional, i.e. explicable-by-content. All processes that are not susceptible to intentional explanation are bundled together under the label 'innate'. That puts acquiring the concept JOHN or CARBURETTOR by rich interactions with John or with actual carburettors on a par with acquiring such concepts via an accidental bump on the head. But Fodor has given us a false dichotomy. It is a familiar point that Fodor's model of concept acquisition as hypothesis testing is too restrictive. The further point is that representational acquisition need not be explicable-by-content at all, but may still be recognisably in the domain of psychological explanation.

This paper has shown that is more than just a theoretical possibility. Connectionist models offer actual examples of that process. In order to see connectionist models as accounting for the development of new representations we have to reject the assumption that development of new representations can always be explained-by-content, an assumption that lies at the heart of Fodor's critique of connectionism, and of his radical concept nativism. This paper argues that we should reject that assumption and embrace the idea that some connectionist models show how new primitive representations can develop in response to the environment, without relying on pre-existing representational resources. Connectionism's answer to the question, 'Over *what* are your representations distributed?' opens up a new portion of logical space for cognitive science. Connectionist representations can be distributed over the objects that figure in an account of their development. In that way, connectionist modelling has provided a deep philosophical insight and an important contribution to theoretical progress in cognitive science: new, non-innate representations can develop in ways that are not explicable-by-content.

Acknowledgements The author would like to thank the follow for generous comments: David Braddon-Mitchell, Steve Butterfill, Nick Chater, Martin Davies, Jerry Fodor, Paul Griffiths, Peter Godfrey-Smith, Richard Holton, Matteo Mameli, David Papineau, Gualtiero Piccinini, Kim Plunkett, Paul Smolensky, Mark Sprevak, Scott Sturgeon; the referees and audiences in Melbourne, Oxford and Sydney.

References

Carey, S. (2009). *The origin of concepts*. Oxford/New York: O.U.P.

Chalmers, D. (1992). Subsymbolic computation and the Chinese room. In J. Dinsmore (Ed.), *The symbolic and connectionist paradigms: Closing the gap*. Hillsdale: Lawrence Erlbaum.

Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy, 95*(1), 5–32.

Churchland, P. M. (2012). *Plato's camera: How the physical brain captures a landscape of abstract universals*. Cambridge, MA: MIT Press.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.

Clark, A. (2001). *Mindware*. Oxford: O.U.P.

Cowie, F. (1999). *What's within?* Oxford: OUP.

Dretske, F. (1988). *Explaining behaviour: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Fodor, J. A. (1975). *The language of thought*. London/Cambridge, MA: MIT Press.

Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations: Philosophical essays on the foundations of cognitive science*. London/Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge, MA: MIT Press.

Fodor, J. A. (1994). *The elm and the expert*. Cambridge, MA: MIT Press/Bradford Books.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. New York: OUP.

Fodor, J. A. (2001). Doing without what's within: Fiona Cowie's critique of nativism. *Mind, 110*, 99–148.

Fodor, J. A. (2004). Distributed representations; enough already. http://www.nyu.edu/gsas/dept/philo/courses/representation/papers/fodordistributed.pdf. Accessed 2 Jun 2014.

Fodor, J. A. (2007). The revenge of the given. In McLaughlin & Cohen (Eds.), *Contemporary debates in philosophy of mind.* Oxford: Blackwell.

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford/New York: O.U.P.

Fodor, J. A., & Lepore, E. (1992). *Holism: A shopper's guide*. Oxford: Blackwell.

Fodor, J. A., & Lepore, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *Journal of Philosophy, 96*(8), 381–403.

Fodor, J. A., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*, 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Johnson, M. H., et al. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*, 1–19.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology, 13*(1), 47–76.

Laurence, S., & Margolis, E. (2002). Radical concept nativism. *Cognition, 86*, 25–55.

Macnamara, J. (1986). *Border dispute: The place of logic in psychology*. Oxford/New York: O.U.P.

Mameli, M. (2008). On innateness: The clutter hypothesis and the cluster hypothesis. *Journal of Philosophy, 55*, 719–736.

Mareschal, D., & Schultz, T. R. (1996). Generative connectionist networks and constructivist cognitive development. *Cognitive Development, 11*, 571–603.

Margolis, E. (1998). How to acquire a concept. *Mind & Language, 13*(3), 347–369.

McLeod, P., Shallice, T., & Plaut, D. C. (2000). Attractor dynamics in word recognition: Converging evidence from errors by normal subjects, dyslexic patients and a connectionist model. *Cognition, 74*, 91–113.

Millikan, R. G. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLEARN: A two process theory of infant face recognition. *Psychological Review, 98*, 164–181.
Papineau, D. (1987). *Reality and representation*. Oxford: Blackwell.
Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology, 10*(3), 209–254.
Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral & Brain Sciences, 20*, 537–596.
Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12*(5), 410–430.
Rolls, E., & Treves, A. (1998). *Neural networks and brain function*. Oxford: OUP.
Rupert, R. D. (1998). On the relationship between naturalistic semantics and individuation criteria for terms in a language of thought. *Synthese, 117*(1), 95–131.
Rupert, R. (2001). Coining terms in the language of thought. *Journal of Philosophy, 98*, 499–530.
Ryder, D. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind & Language, 19*(2), 211–240.
Samuels, R. (2002). Nativism in cognitive science. *Mind & Language, 17*, 233–265.
Sejnowski, T., & Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1*, 145–168.
Shea, N. (2007a). Consumers need information: Supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research, 75*(2), 404–435.
Shea, N. (2007b). Content and its vehicles in connectionist systems. *Mind & Language, 22*(3), 246–269.
Shea, N. (2011). New concepts can be learned, review essay on Susan Carey. The Origin of Concepts. *Biology & Philosophy, 26*, 129–139.
Shea, N. (2012a). Genetic representation explains the cluster of innateness-related properties. *Mind & Language, 27*(4), 466–493.
Shea, N. (2012b). New thinking, innateness and inherited representation. *Philosophical Transactions of the Royal Society B, 367*, 2234–2244.
Strevens, M. (2012). Theoretical terms without analytic truths. *Philosophical Studies, 160*(1), 167–190.
Tiffany, E. (1999). Semantics San Diego style. *Journal of Philosophy, 96*, 416–429.

Chapter 15
Toward a Theory of Intelligent Complex Systems: From Symbolic AI to Embodied and Evolutionary AI

Klaus Mainzer

Abstract In the twentieth century, AI (artificial Intelligence) arose along with Turing's theory of computability. AI-research was focused on using symbolic representations in computer programs to model human cognitive abilities. The final goal was a complete symbolic representation of human intelligence in the sense of Turing's AI-test. Actually, human intelligence is only a special example of problem solving abilities which have evolved during biological evolution. In embodied AI and robotics, the emergence of intelligence is explained by bodily behavior and interaction with the environment. But, intelligence is not reserved to single organisms and brains. In a technical coevolution, computational networks grow together with technical and societal infrastructures generating automated and intelligent activities of cyberphysical systems. The article argues for a unified theory of intelligent complex systems.

Keywords Classical AI • Embodiment • Networks • Intelligent complex systems

15.1 Symbolic and Embodied Artificial Intelligence

15.1.1 What Are Intelligent Systems?

Intelligent systems are a subclass of information systems which can be found in nature, technology, and society. There are different examples of intelligent systems – animals, primates and humans, populations and societies, computers and robots, information and communication networks. They all are distinguished by different kinds and degrees of intelligence, sometimes in interaction and in dependence of humans. But with increasing autonomy of agents and robots populations and

K. Mainzer (✉)
Technical University Munich, Munich, Germany
e-mail: mainzer@cvl-a.tum.de

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_15

with self-organization of information and communication networks, we observe the technical development of intelligent artificial systems surpassing natural evolution of organisms and populations. We define that intelligence only means an ability to solve problems. The degree of intelligence depends on the class of problems to be solved. In this case, intelligence is not reserved to humans. A tick is intelligent with respect to a particular class of problem solving (e.g., finding the blood of a host organism) as well as a computer program of pattern recognition. Of course, there are hybrid systems with several intelligent abilities like, e.g., humans and primates. But, these are only differences of complexity degrees, not in principle. In a rigorous way, complexity degrees can be measured by the algorithmic tools of computational complexity, i. e. time or size of the procedures to solve a problem. The distinction of natural and artificial systems is only justified by the fact that "artificial" systems are once initiated by human technology. But, in the future, originally "artificial" systems may reproduce and organize themselves in an automated evolution (Mainzer 2003, 2010).

In the past, intelligent systems have only been developed during biological evolution by more or less random conditions. During the factual evolution on Earth, many species occurred according to the laws of evolution, but many of them disappeared again because of less adaptability to changing living conditions. The laws of evolution, e.g., self-replication, selection, and mutation, are independent of particular molecular substances. In the past, they were only applied to particular biomolecules on Earth. But, in cellular automata, they are virtually simulated, and in synthetic biology, they are already applied to new constructed chemical building blocks which were unknown in the chemistry of life on Earth. Thus, the criteria of life do not depend on a particular species on Earth. On this line, we also argue that the definition of intelligence does not depend on the abilities of a particular species on Earth. Thus, the traditional Turing test of artificial intelligence is less helpful, because it relates machine intelligence to human intelligence. Furthermore, in the universe, new forms of life and intelligence are highly probable, even though still unknown.

15.1.2 Symbolic AI

Computational systems were historically constructed on the background of Turing's theory of computability. In Turing's functionalism, the hardware of a computer is related to the wetware of human brain. The mind is understood as the software of a computer. Turing argued: If human mind is computable, it can be represented by a Turing program (Church's thesis) which can be computed by a universal Turing machine, i.e. technically by a general purpose computer. Even if people do not believe in Turing's strong AI (Artificial intelligence)-thesis, they often claim classical computational cognitivism in the following sense: Computational processes operate on symbolic representations referring to situations in the outside world. These formal representations should obey Tarski's correspondence theory

of truth (Tarski 1935): Imagine a real world situation X1 (e.g., some boxes on a table) which is encoded by a symbolic representation A1 = encode(X1) (e.g., a description of the boxes on the table). If the symbolic representation A1 is decoded, then we get the real world situation X1 as its meaning, i.e. decode(A1) = X1. A real-world operation T (e.g., a manipulation of the boxes on the table by hand) should produce the same real-world result A2, whether performed in the real world or on the symbolic representation:

$$\text{decode}(\text{encode}(\text{T})(\text{encode}(\text{X1}))) = \text{T}(\text{X1}) = \text{X2}.$$

Thus, there is an isomorphism between the outside situation and its formal representation. As the symbolic operations are completely determined by algorithms, the real-world processes are assumed to be completely controlled. Therefore, classical robotics operates with completely determined control mechanisms.

Symbolic representations with ontologies, categories, frames, and scripts of expert systems work along this line. But, they are restricted to a specialized knowledge base without the background knowledge of a human expert. Human experts do not rely on explicit (declarative) rule-based representations only, but also on intuition and implicit (procedural) knowledge (Dreyfus 1979). Further on, our understanding depends on situations. The situatedness of representations is a severe problem of informatics. A robot needs a complete symbolic representation of a situation which must be updated if the robot's position is changed. Imagine that it circles around a table with a ball and a cup on it. A formal representation in a computer language may be ON(TABLE,BALL), ON(TABLE,CUP), BEHIND(CUP,BALL), et alt. Depending on the robot's position relative to the arrangement, the cup is sometimes behind the ball or not. So, the formal representation BEHIND(CUP,BALL) must always be updated in changing positions.

15.1.3 Embodied AI

But, how can the robot prevent incomplete knowledge? How can it distinguish between reality and its relative perspective? Situated agents like human beings need no symbolic representations and updating. They look, talk, and interact bodily, for example, by pointing to things. Even rational acting in sudden situations does not depend on symbolic representations and logical inferences, but on bodily interactions with a situation (for example, looking, feeling, reacting).

Thus, we distinguish formal and embodied acting in games with more or less similarity to real life: Chess is a formal game with complete representations, precisely defined states, board positions, and formal operations. Soccer is a non-formal game with skills depending on bodily interactions, without complete representations of situations and operations which are never exactly identical. According to the French philosopher Merleau-Ponty, intentional human skills do not need any symbolic representation, but they are trained, learnt, and embodied by the organism (Merleau-Ponty 1962; Dreyfus 1982). An athlete like a pole-vaulter cannot repeat her

successful jump like a machine generating the same product. The embodied mind is no mystery. Modern biology, neural, and cognitive science give many insights into its origin during the evolution of life.

15.2 Embodied Mind and Brain Dynamics

The coordination of the complex cellular and organic interactions in an organism needs a new kind of self-organizing controlling. Their development was made possible by the evolution of nervous systems that also enabled organisms to adapt to changing living conditions and to learn bodily from experiences with its environment. We call it the emergence of the embodied mind (Mainzer 2009). The hierarchy of anatomical organizations varies over different scales of magnitude, from molecular dimensions to that of the entire central nervous system (CNS). The research perspectives on these hierarchical levels may concern questions, for example, of how signals are integrated in dendrites, how neurons interact in a network, how networks interact in a system like vision, how systems interact in the CNS, or how the CNS interact with its environment.

15.2.1 Mathematical Modeling of Brain Dynamics

In the complex systems approach, the microscopic level of interacting neurons can be modeled by coupled differential equations modelling the transmission of nerve impulses by each neuron. The Hodgekin-Huxley equation is an example of a nonlinear reaction diffusion equation of a travelling wave of action potentials which give a precise prediction of the speed and shape of the nerve impulse of electric voltage. In general, nerve impulses emerge as new dynamical entities like the concentric waves in BZ-reactions or fluid patterns in non-equilibrium dynamics.

But, local activity of a single nerve impulse is not sufficient to understand the complex brain dynamics and the emergence of cognitive and mental abilities. The brain with its more than 10^{11} neurons can be considered a huge nonlinear lattice, where any two points (neurons) can interact with neural impulses. How can we bridge the gap between the neurophysiology of local neural activities and the psychology of mental states? A single neuron can neither think nor feel, but only fire or not fire. They are the "atoms" of the complex neural dynamics.

In his famous book "The organization of Behavior" (1949), Donald Hebb suggested that learning must be understood as a kind of self-organization in a complex brain model. As in the evolution of living organisms, the belief in organizing "demons" could be dropped and replaced by the self-organizing procedures of the self-organizing procedures of the complex systems approach. Historically, it was the first explicit statement of the physiological learning rule for synaptic modification.

Hebb used the word "connectionism" in the context of a complex brain model. He introduced the concept of the Hebbian synapse where the connection between two neurons should be strengthened if both neurons fired at the same time (Hebb 1949, 50):

> When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

Hebb's statement is not a mathematically precise model. But, later on, it was used to introduce Hebb-like rules tending to sharpen up a neuron's predisposition "without a teacher" from outside. For example, a simple mathematical version of Hebb's rule demands that the change Δw_{BA} of a weight w_{BA} between a neuron A projecting to neuron B is proportional to the average firing rate ν_{A} of A and ν_{B} of B, i.e., $\Delta w_{\mathrm{BA}} = \varepsilon\, \nu_{\mathrm{B}} \nu_{\mathrm{A}}$ with constant ε. In 1949, the "Hebbian synapse" could only be a hypothetical entity. Nowadays, its neurophysiological existence is empirically confirmed.

On the macroscopic level, Hebb-like interacting neurons generate a cell assembly with a certain macrodynamics (Haken 1996). Mental activities are correlated with cell assemblies of synchronously firing cells. For example, a synchronously firing cell-assembly represents a plant perceptually which is not only the sum of its perceived pixels, but characterized by some typical macroscopic features like form, background or foreground. On the next level, cell assemblies of several perceptions interact in a complex scenario. In this case, each cell-assembly is a firing unit, generating a cell assembly of cell assemblies whose macrodynamics is characterized by some order parameters. The order parameters may represent similar properties of the perceived objects.

There is no "mother neuron" which can feel, think, or at least coordinate the appropriate neurons. The binding problem of pixels and features in a perception is explained by cell assemblies of synchronously firing neurons dominated by learnt attractors of brain dynamics. The binding problem asked: How can the perception of entire objects be conceived without decay into millions of unconnected pixels and signals of firing neurons? Wolf Singer (Singer 1994) and others could confirm Donald Hebb's concept of synchronously firing neurons by observations and measurements.

In this way, we get a hierarchy of emerging levels of cognition, starting with the microdynamics of firing neurons. The dynamics of each level is assumed to be characterized by differential equations. For example, on the first level of macrodynamics, their solutions determine a visual perception. On the following level, the observer becomes conscious of the perception. Then the cell assembly of perception is connected with the neural area that is responsible for states of consciousness. In a next step, a conscious perception can be the goal of planning activities. In this case, cell assemblies of cell assemblies are connected with neural areas in the planning cortex, and so on. They are represented by coupled nonlinear equations with firing rates of corresponding cell assemblies. Even high-level concepts like self-consciousness can be explained by self-reflections of self-reflections, connected

with a personal memory which is represented in corresponding cell assemblies of the brain. Brain states emerge, persist for a small fraction of time, then disappear and are replaced by other states. It is the flexibility and creativeness of this process that makes a brain so successful in animals for their adaption to rapidly changing and unpredictable environments.

From a mathematical point of view, the interactions of n assemblies are described by a system of n coupled differential equations depending on their specific common firing rates F_j ($j = 1, 2, \ldots, n$) with

$$\frac{dF_1}{dt} = F_1\,(1 - F_1) - \alpha\, F_2 - \alpha\, F_3 - \ldots \alpha\, F_n,$$

$$\frac{dF_2}{dt} = -\alpha\, F_1 + F_2\,(1 - F_2) - \alpha\, F_3 - \cdots - \alpha\, F_n,$$

$$\vdots$$

$$\frac{dFn}{dt} = -\alpha\, F_1 - \alpha\, F_2 - \alpha\, F_3 - \cdots + F_n\,(1 - F_n)\,,$$

where α is a parameter that is positive for inhibitory interactions among assemblies (Scott 2003).

Cell assemblies behave like individual neurons. Thus, an assembly of randomly interconnected neurons has a threshold firing level for the onset of global activity. If this level is not attained, the assembly will not ignite, falling back to a quiescent state. If the threshold level is exceeded, firing activity of an assembly will rise rapidly to a maximum level. These two conditions ensure that assemblies of neurons can form assemblies of assemblies. Assemblies emerge from the nonlinear interactions of individual neurons, assemblies of assemblies emerge from the nonlinear interaction of assemblies. Repeated several times, one gets the model of the brain as an emergent dynamic hierarchy.

15.2.2 Emergence of Intelligence in the Brain

In brain research, it is assumed that all mental states are correlated to cell assemblies. The corresponding cell assemblies must empirically be identified by observational and measuring instruments. In brain reading, for example, active cell assemblies correlated with words and corresponding objects can be identified. A single neuron is not decisive and may differ among different persons. There are typical distribution patters with fuzzy shapes which are represented in computer simulations. Brain research is still far from observing the activities of each neuron in a brain. Nevertheless, the formal hierarchical scheme of dynamics, at least, allows an explaining model of complex mental states like, for instance, consciousness. In this model, conscious states mean that persons are aware of their activities. Self-awareness is realized by additional brain areas monitoring the neural correlates of

these human activities (e.g., perceptions, feeling, or thinking). This is a question of empirical tests, not of arm-chaired reflection. For example, in medicine, physicians need clear criteria to determine different degrees of consciousness as mental states of patients, depending on states of their brain.

Thus, we aim at clear working hypotheses for certain applications and not at a "complete" understanding what "consciousness" means per se. Besides medicine, the assumption of different degrees of self-awareness opens new perspectives of technical applications. Robots with a certain degree of self-awareness can be realized by self-monitoring and self-control which are useful for self-protection and cooperation in robot teams. In technical terms, these robots have internal representations of their own body and states. They can also be equipped with internal representations of other robots or humans which can be changed and adapted by learning processes. Thus, they have their own "theory of mind" with perspectives of first and second person. In this sense, even consciousness is no mysterious event, but observable, measurable, and explainable in appropriate research frameworks. The formal hierarchical model offers the opportunity to build corresponding circuits and technical equipment for technical brains and robots with these abilities.

Traditional terms "intelligence", "mind", "consciousness" etc. are historically overloaded with many meanings depending on different point of views, experience and historical positions (Chalmers 2010). Therefore, it depends on our working definitions. Concerning intelligence, a working definition is suggested which does not depend on "human intelligence" (in the sense of Turing's AI-test). A system is called "intelligent" depending on its ability to solve problems. Simple technical systems (e.g., a chip) can also have certain degrees of "intelligence", because they can solve certain problems. Thus, in philosophical terms, this position sympathizes with the pluralism of Leibniz who distinguished degrees of intelligence in nature instead of Descartes' dualism who believed in a "substance" called "intelligent mind" which was reserved to human beings. Following our scaling of complexity degrees, there are already many intelligent functions of, e.g., robots. According to our working definition, it is remarkable that intelligent problem solving does not necessarily need "consciousness". For example, effective pattern recognition with technical neural networks does not need consciousness which is typically connected with human recognition. It is amazing how effective technical systems solve complex problems without consciousness, sometimes even better.

Obviously, patterns of cell assemblies in the brain are not identical with our perceptions, feeling, and thinking. But, it is well confirmed in modern brain research that neural patterns of firing cells are correlated with mental states. These correlates can be mathematically defined and modeled in state and parameter spaces with associated dynamical systems which allow us to test our models (Mainzer and Chua 2013). With the technology of brain reading, an analysis of cell assemblies was used to extract correlates of what is represented (e.g., pictures, words, phrases): Of course, there are only the first steps of research, but it seems to be possible at least in principle. Concerning computer science, semantics in the sense of correlated contexts is technically realized in first steps and to certain degrees in a restricted and well-defined sense (cf. 2.3).

Motor, cognitive, and mental abilities are stored in synaptic connections of cell assemblies. A hard core of synaptic network is already wired, when a mammal brain is born. But many synaptic connections are generated during growth, experience and learning phase of mammals. Firing states of neurons with repeated action potentials enforce synaptic connections. Thus, during a learning phase, a cell assembly of simultaneously firing neurons creates a synaptic network storing the learnt information. Learning phases can be modeled technically by learning algorithms (Mainzer 2007). As we all know, the learnt information can be forgotten, when learning is not repeated and the synaptic connections decay. Thus, on the micro level, brain dynamics is determined by billions of firing and not firing neurons, and, on the macro level, by emerging and changing cell assemblies of neural networks coding different neural information.

The efficiency of neural networks depends on their number of hierarchical layers. They enable the brain to connect different neural states of, e.g., visual, haptic, and auditory information. But, there are also layers monitoring perceptual procedures and generating visual consciousness: A person is aware and knows that she perceives something. Even our emotions depend on specified neural networks which are connected with all kinds of brain activity. It is a challenge of brain research to identify the involved layers and networks of the brain during all kinds of mental and cognitive activities.

15.2.3 Emergence of Semantic Meaning in the Brain

Semantic understanding is made possible by hierarchical layers and learning procedures of brain dynamics. In formal semantics, a formal language gets its meaning by mapping its formal expressions onto expressions of another formal language. In computer science, several layers of symbolic languages are also used to generate semantic meaning and digital procedures. Natural languages of users refer to an advanced computer language. In a next step, a kind of behavioral language describes the functions that are intended to be executed in the system. They relate their arguments typically like mathematical functions. Then, an architectural description is needed that describes the available resources, the communication and the control. The relation between the behavioral description and the architecture is done by binding elementary functions to resources and scheduling them in time (a task that can be done by a compiler or an assembler). Finally, the architecture gets realized by a register transfer layer which is close to digital circuits that in turn are realized by electrical components, voltages and currents. Each such layer has its own descriptive language, dynamics and syntactic rules. Thus, formal semantics is defined by interaction between layers in a rigorous logical-mathematical sense: Each layer must “understand” the meaning of their instructions and transfer rules. The digital circuit provides the semantics for the electrical circuit (vice versa, the circuit layer provides timing information to the digital), the register transfer to the digital layer, the architecture to the register transfer layer and the behavioral

description sits on top of that all, at least five layers of semantic abstractions, each expressed in a dedicated formal syntax. It is only because engineers can make such a strong taxonomy and abstraction of the semantic layers, that a digital system is understandable for the human user and the digital machine. The communication between man and machine is realized in an intermediate transfer process between layers.

Some people still believe that "meaning" is something mysterious which can only generate by humans, and "stupid" computers only work on the syntactical level. Arguing on this line, "meaning" only comes into the machine on the most abstract layer by human programmers and engineers. Since more than 100 years, we can already identify the areas of the brain which are involved in semantic language processing. The Brocca-area (in the left frontal lobe) regulates syntax, Wernicke-area (in the left temple) is responsible for semantic processing. These areas are connected by complex fibers of nerves with two separated circuits for syntactic and semantic processes. The efficiency of information processing is guaranteed by an insulating myelin layer. In an adult brain, we can distinguish three bundles of nervous fibers: The first bundle connects the areas which enable syntactic processing (e.g., area 44) in a circuit. A first bundle connects the areas which are responsible for meaning of words and sentences (e.g., area 45). A third bundle connects the premotor cortex with the auditory cortex and supports repeating of heard words. Together with the right hemisphere for melodic aspects of language, we get a highly efficient network of human language processing.

In the end, the brain must be totally scanned and modeled from its single neurons, synapses, and action potentials to cell assemblies, networks, and layers, in order to model the whole dynamics on the micro- and macrolevel. The "machine level" is already well known and described by emerging action potentials. Their patterns could be identified in parameter spaces of the Hodgkin-Huxley equations. Their emergence can be explained and predicted by their local activity with rigorous mathematical methods (Mainzer and Chua 2013). They are the origin of all kind of brain dynamics with attractors of neural states correlated with human cognitive and intelligent activities (Freeman 2004). But like conventional computers, the brain alone is not able to generate meaning. Meaning of languages emerges from interaction with different environments in different linguistic communities. In the case of human beings, complex bodily abilities of sensor perception, feeling, and imagination are involved as well as societal experiences. It will be a challenge of brain research, cognitive science, and linguistics to integrate their research, in order to capture these complex activities. Embodied robotics can use their insights to model them in technical systems.

But there is another way of technology to overcome the complexity of human language processing. Compared with human brains, technical systems may be restricted, but they are sometimes much more effective with their specific solutions of cognitive and intelligent tasks. In computer science, semantic webs and i-phones can already understand questions to some extent and even answer in natural languages. The technology of applied (speech analysis) algorithms may be different from biological procedures which were developed during evolution. But, they solve

the problem to some degree with their computer power, high speed, parallelism and storage which can be improved in the future.

These procedures can be illustrated by automated translations of two languages with Big Data algorithms (Mainzer 2014). A human translator must know grammar, vocabulary and the filigree meaning of both languages. The reason is the big number of words and phrases with multi-meaning depending on different contexts. Thus a human translator must not only master all the nuances of both languages, but also the contents of texts. This task can be managed by statistical methods on a very high level. It is not necessary to speak or to understand both languages. Further on, you do not need a linguistic expert who, together with a programmer, feed a computer with linguistic knowledge or rules. You only need a mass of data in a pool with translated texts from a source language into a target language. The Internet is an example of such a powerful store. In the meantime, nearly every group of words is translated by anyone and anywhere for several times. Parallel texts are the basis of this kind of translations. The probability that a translation is close to a text increases with the frequency that a group of words in the data pool is translated in a certain context in a certain way. The context of words can be determined quantitatively by a computer with statistical measure numbers. Thus, from a technical point of view, we must not understand what "understanding" means in all its filigree meaning. May be that cognitive science and brain research will be successful someday to do that. In the end, we are already mastering linguistic challenges by powerful data bases and algorithms in a better way than human linguistic experts ever did.

15.3 Embodied and Cognitive Robotics

Embodied computing applies the principles of evolution and life to technical systems (Mainzer 2009). The dominating principles in the complex world of evolution are self-organization and self-control. How can they be realized in technical systems? In many cases, there is no finite program, in order to forecast the development of complex systems. In general, there are three reasons for computational limits of system dynamics: (1) A system may be undecidable in a strict logical sense. (2) Further on, a system can be deterministic, but nonlinear and chaotic. In this case, the system depends sensitively on tiny changes of initial data in the sense of the butterfly effect. Long-term forecasting is restricted, and the computational costs of forecasting increase exponentially after some few steps of future predictions. (3) Finally, a system can be stochastic and nonlinear. In this case, pattern emergence can only be predicted probabilistically.

In complex dynamical systems of organisms, monitoring and controlling are realized on hierarchical levels. Thus, we must study the nonlinear dynamics of these systems in experimental situations, in order to find appropriate models and to prevent undesired emergent behavior as possible attractors. From the point of view of systems science, the challenge of embodied robotics is controlled emergence.

In the research project "Cognition in Technical Systems" (CoTeSys 2006–2012), cognitive and life sciences, information processing and mathematical sciences, engineering and robotics work systematically together to explore cognition for technical systems. Robotic agents cannot be fully programmed for every application (Shuji Kajita 2007). The program learns from experience where to stand when taking a glass out of a cupboard, how to best grab particular kitchen utensils, where to look for particular cutlery, et alt. This requires the control system to know the parameters of control routines and to have models for how the parameters change the behavior. The sensor data of a robot's environment, which is the robot's "experience", are stored in a relational database system, the robot's "memory". According to the paradigm of probabilistic robotics (Thrun et al. 2005), the data in the database together with causal structure on domain relations imply a joint probability distribution over relations in the activity domain. This distribution is applied in Markov logic, which allows inferring the conditional probability of logical (first order) statements. In short: A robot can estimate the environmental situation probabilistically.

According to the paradigm of complex dynamical systems, a robot can be described at different levels, in which global properties at one level emerge from the interaction of a number of simple elements at lower levels. Global properties are emergent in the sense that they result from nothing else but local interactions among the elements. They cannot be predicted or inferred from knowledge of the elements or of the rules by which the elements locally interact, given the high nonlinearity of these interactions.

Simple examples of embodied robotics are reactive robots. They are controlled by simple neural networks, for example, fully connected perceptrons without internal layers and without any kind of internal organization. Nevertheless, these robots can display not only simple behaviors, such as obstacle avoidance, but also behaviors capable of solving complex problems involving perceptual aliasing, sensory ambiguity, and sequential organization of sub-behaviors. The question arises how far we can go with reactive sensory-motor coordination.

Not only "low level" motor abilities, but also "high level" cognition (for example, categorization) can emerge from complex bodily interaction with an environment by sensory-motor coordination without internal representation in symbolic syntax. We call it "embodied cognition": Developmental psychology shows that an infant learns to associate and categorize objects and to build up concepts by touching, grasping, manipulating, feeling, tasting, hearing, and looking at things, before it is able to use syntactical rules of linguistic symbols. The reason is that language acquisition follows a biological program (Friederici 2006): In the brains of babies, the three bundles of nervous fibers which are responsible for syntactic, semantic, and auditory language processing (cf. 2.3) do already exist. Only the circuits of semantic and auditory processing are developed and can be visualized in the brains of babies. But the third bundle of fibers which enables the application of syntactic symbolic rules is not developed before the age of three years. The neural networks need more than eight years to realize their final efficiency of symbolic syntax processing.

These observations motivate embodied robotics. Symbolic representation refers to linguistic signs in natural languages or formal terms in programming languages. A conventional computer only works according to the symbolic instructions of a computer language. But embodied robots react and act by sensor inputs and haptic actuators with their environment. In an analogue way, the categories of human infants start with impressions and feelings based on sensor circuits which will be connected with linguistic representations in later stages of development. We have an innate disposition to construct and apply conceptual schemes and tools at a certain stage of development (Bellman 2005; Mainzer 2008a, b). In embodied robotics, robots are equipped with neural networks to recognize correlations and patterns of sensor data for their orientation. They can be connected with frames and schemes of AI which support a robot to categorize objects and situations. So, on a higher level, symbolic AI comes in like natural languages and formal representations in later stages of human development. The whole Internet can be used as big "memory" of robots which overcomes the limitations of human brains. We are far away from capturing all cognitive human abilities in technical systems. But, in some special fields, there are, again, technical strategies which are much more effective than our evolutionary equipment (e.g., pattern recognition of Big Data algorithms with the Internet).

15.4 Technical Coevolution of Intelligent Cyberphysical Systems

Technical neural networks are complex systems of firing and non-firing neurons with topologies like living brains. There is no central processor (mother cell), but a self-organizing information flow in cell-assemblies according to rules of synaptic interaction. The cells are amplifiers of low input signals generating complex patterns of cellular behavior. There are different architectures with one synaptic layer or more synaptic layers with hidden units for feed-forward flow, or feedback loops with back-propagation flow. Learning algorithms change the synaptic weights, in order to realize the synaptic plasticity of living brains. The dynamics of neural nets can be modeled in phase spaces of synaptic weights with trajectories converging to attractors which represent prototypes of patterns. Neural networks are already applied in cognitive robots. A simple robot with diverse sensors (e.g., proximity, light, collision) and motor equipment can generate intelligent behavior by a self-organizing neural network. Intelligent behavior can also be generated by interaction of robots in robot societies or in the internet of things.

15.4.1 Artificial Intelligence of Robot Societies

A robot society is a group of robots which has the ability to communicate, interact, and to solve problems jointly. By that, a robot society can generate intelligent

behavior like interacting ants of a population or interacting neurons of brains. A society is defined by its information and control structure which make possible common task planning and execution. In this case, a robot is a locally active agent driven by a battery and low input signals which are amplified and transformed into complex patterns of behavior.

Most of the autonomous mobile robots are operating in neither stabile nor structured environments (Bekey 2005). Therefore, a major trend in robotics is going towards multi-robot systems (Balch and Parker 2002). In many cases, the decomposing of a complex task into parallel subtasks is a possibility to speed up the performance (Mataric et al. 2003). Sometimes, several robots work with the same subtask, increasing the redundancy of the system. Furthermore, there can be tasks where a successful completion of a task requires close cooperation among the robots. Such case is, for example, the carrying of a large object together. It requires some sort of interaction between robots, whether is a direct communication or some sort of indirect communication through sensing the forces in the object to be transported.

This kind of task as well as many other tasks normally related to multi-robot systems has clear analogy to biological systems (Wilson 2000). A group of ants solve their behavioral problems through sensing the forces and torque in the object. Based on this information they change the direction of forces accordingly or needed some ants change the position of their hold. Numerous similar examples can be found from nature. Tests by evolution during millions of years are proven to be feasible in dynamic and hostile environments and can thus provide valuable information and inspiration for similar type of engineering tasks.

A profound challenge of interdisciplinary importance is the question how can intelligence emerge and evolve as a novel property in groups of social animals and robots. According to our definition of intelligence, robot societies have intelligence with a certain degree, if they can solve problems with a certain degree of complexity. The question can be solved by focusing the attention on the very early stages of the emergence and evolution of simple technical artefacts. Therefore, one should start by building an artificial society of embodied agents as real robots, creating an environment or artificial ecosystem and appropriate primitive behaviors for those robots, then free running the artificial society. Even with small populations of simple robots, a large number of interactions between robots can be generated (Brooks 1999; Braitenberg and Radermacher 2007; Pfeifer and Scheier 2001). The inherent heterogeneities of real robots, and the noise and uncertainty of the real world, increase the space of possibilities and the scope for unexpected emergence in the interactions between robots.

The goal is to create the conditions in which proto-culture can emerge in a robot society. Robots can copy each other's behaviors and select which behaviors to copy. Behaviors will mutate because of the noise and uncertainty in the real robots' sensors and actuators. Successful types of behavior will undergo multiple cycles of copying (heredity), selection and variation (mutation). With evolutionary time, a genetic algorithm process to grow and evolve the robots' controllers so that the emerging patterns of behavior become hard-wired into the robots' controllers (Nolfi and Floreano 2001).

15.4.2 Social Intelligence of Sociotechnical Systems

Social networks of more or less autonomous robots are only one possible development in a general trend of future technology. In a technical co-evolution, global information and communication networks are emerging with surprising similarity to self-organizing neural networks of the human brain. The increasing complexity of the World Wide Web (www) needs intelligent strategies of information retrieval and learning algorithms simulating the synaptic plasticity of a brain (Berners-Lee 1999). The Internet links computers and other telecommunication devices. At the router level, the nodes are the routers, and the edges are their physical connections. At the interdomain level, each domain of hundreds of routers is represented by a single node with at least one route as connection with other nodes. At both levels, the degree distribution follows a power law of scale-free network which can be compared with the networks in systems biology. Measurements of the clustering coefficient deliver values differing from random networks and significant clusters. The average paths at the domain level and the router level indicate the small-world property.

In the future, global information networks will grow together with societal infrastructure in cyberphysical systems (acatech 2011). Current examples are complex smart grids of energy. Many energy providers of central generators and decentralized renewable energy resources lead to power delivery networks with increasing complexity. Smart grids mean the integration of the power delivery infrastructure with a unified communication and control network, in order to provide the right information to the right entity at the right time to take the right action. It is a complex information, supply and delivery system, minimizing losses, self-healing and self-organizing.

Smart grids are complex organizations of networks regulating, distributing, storing, and generating electrical power. Their structure and dynamics have surprising similarity with complex protein networks in systems biology regulating the energy supply of a cell. The intelligence of smart grids increases with their ability of self-organizing information processing for optimal energy supply. In communication networks, appropriate prices of optimal energy supply could be automatically negotiated by virtual agents. In smart grids, the energy system grows together with information and communication technology in a kind of symbiosis.

A well-known problem with wind mills and solar cells is the unpredictability of production depending on changing weather conditions. In intelligent networks, the need can be locally satisfied by virtual negotiations. A model assumes the following rules and conditions of negotiating virtual agents (Wedde et al. 2008):

1. The need for renewable energy can be satisfied either in a local regional subnet or between subnets. Reserve capacity is used only in exceptional cases.
2. Energy must be adjusted between different voltage levels or different groups of balance on the same level.
3. Producers are also consumers and vice versa.

4. Negotiations on local energy supply are automatically performed by agents of producers and agents of consumers. They are coordinated by balance group managers working parallel and synchronized in time on each level.
5. In the model, the negotiations start in periods of 0.5 s. The negotiations as well as the distribution of negotiated energy are expected to be finished before the end of each period. Bids and offers arriving in the meantime are negotiated in the next period.
6. At the beginning of each period, each client decides whether he/she takes part as producer or consumer or not. He/she decides with respect to the current difference between the states of demand and production.
7. Bids and offers occur in frameworks of prices with respect to amortization and maintenance. In the model, there are no long-range contracts or discounts for big and future acquisitions which can occur in reality.

The algorithm of negotiation assumes a framework of prices for each level of negotiation. Each balance group manager on each level accomplishes a cycle of coordination of 10 turns. Each turn takes 1 ms. After each turn the balance managers test in parallel whether bids and offers are sufficiently similar. If they are sufficiently similar, a contract between the partners is concluded. A fixed amount is added until the stock or demand is spent. The negotiation strategies of a client are given by an opening bid, an opening offer, and parameters of priority and strategy. After n turns, the unsatisfied agents adapt their bids and offers with respect to an exponential law of behavior which is useful to realize a fast convergence between bids and offers. The negotiated price is the arithmetic mean between similar values. Unsatisfied clients are passed on to the next level of negotiation. On this level, the framework of prices is reduced to a constant relation. The needs and interests of finally unsatisfied clients are satisfied by a central reserve capacity (but with very bad prices).

Short term fluctuations of consumption in the ms to min interval, which are effected by sudden and unpredicted local or regional causes, are not only observed as perturbations in households, but they can endanger the stability of large transport networks. In our model, these critical situations are avoided by the activation of agents after each cycle of negotiation. It is assumed that many electrical appliances (e.g., refrigerator, boiler) can temporarily work without power or with a battery. In these cases, reserve energy can be used for other purposes. The reserve energy is more competitive than the traditional one, because of low costs of transport and storage in the network. Additionally, the balance managers act on each level in parallel in shortest time.

Smart grids with integrated communication systems accomplish a dynamical regulation of energy supply. They are examples of large and complex real-time systems according to the principles of cyber-physical systems (Lee 2008). Traditionally, reserve energy which is used to balance peaks of consumption or voltage drops is stored by large power plants. The main problem of changing to renewable energies is the great number of constraints depending on questions of functionality as well a security, reliability, temporary availability, tolerance of failures, and adaptability. Cyber-physical systems with local and bottom-up structures are the best answer to

the increasing complexity of supply and communication systems (Cyber-Physical Systems 2008). In a technical co-evolution mankind is growing together with these technical infrastructures. Their collective abilities emerge like swarm intelligence of populations in evolution which are sometimes called "superorganisms".

Increasing computational power and acceleration of communication need improved consumption of energy, better batteries, miniaturization of appliances, and refinement of display and sensor technology (Weiser 1991; Hansmann 2001). Under these conditions, intelligent functions can be distributed in a complex network with many multimedia terminals. Together with satellite technology and global positioning systems (GPS), electronically connected societies are transformed into cyberphysical systems. They are a kind of symbiosis of man, society, and machine. Communication is not only realized between human partners with natural languages, but with the things of this world. Cyberphysical systems also mean a transformation into an Internet of Things. Things in the Internet become locally active agents.

15.5 Unified Theory of Intelligent Complex Systems

Intelligent systems are subclasses of complex dynamical systems. Different disciplines are growing together in a unified theory of complex networks: systems and synthetic biology, brain and cognition research, software and hardware engineering, robotics, information and communication networking, construction of cyberphysical systems and living infrastructures. The common laws of this unified theory of complex networks are the theory of complex dynamical systems. Applications are self-organizing gene and protein networks, cellular organisms, agents and robots population, cyberphysical systems and communication networks. They all are typical examples of networks with locally active agents amplifying and transforming low energy and input signals into new patterns, structures, and behavior (Mainzer and Chua 2013). They are intelligent with a certain degree, if they solve problems with a certain degree of complexity. The unified theory in these fields is not yet completely accomplished, and, sometimes, we only know certain aspects and main features. But, at least in the natural sciences, it is not unusual to work successfully with incomplete theories: in elementary particle physics, the standard theory of unified forces has still many open questions, but it is nevertheless successfully applied to measure and solve problems.

Complex networks without recognizable patterns are described as random graphs (Albert and Barabási 2002). They start with N nodes and connect pair of nodes with a probability p. With this procedure, we get a graph with $pN(N - 1)/2$ randomly distributed edges. In many networks, there are relatively short connecting paths between two edges ("small worlds"). In complex molecular, cellular, social or technical networks, there are often patterns of, e.g., molecular clusters, cellular assemblies, social groups, circuit diagrams or wireless connections.

During biological, social or technical evolution, they were mainly selected as robust structures which are assigned with biological, social or technical functions of regulation, control, logistics or communication. These clusters are also typical examples of locally active centers emerging from random networks and generating highly structured patterns. Protein networks with their integrated control, logistics, and information systems are examples of cyberphysical systems which are currently developing in social and technical networks of society.

During evolution, cyberphysical systems in nature at first emerged as subcellular logistics, control and information systems in complex gene and protein networks. Neurons were developed as specialized cells of information systems on the basis of neurochemical signal processing. Their networks are models of brains as well as ant populations or human societies. During evolution, effective information and logistic procedures were developed without symbolic representation in computer models. Subcellular, cellular, and neural self-organization generated the appropriate networks. They are equivalent to complex systems which are modeled by nonlinear differential equations. Dynamical systems and their differential equations can be simulated, at least in principle, by computer models. Examples are cellular automata or neural networks with deterministic and stochastic algorithms.

At this point, a deep equivalence of evolutionary, mathematical, and technical procedures become obvious, leading to an extension of Church's famous thesis of computability: not only effective procedures with mathematical symbols can be simulated with computers in the sense of a universal Turing machine, but also atomic, molecular, and cellular coded effective procedures of nature. If the extended Church's thesis holds, then new avenues of computational technologies are opened: all dynamic effective procedures can be modeled on a universal computer (Mainzer and Chua 2011). The symbolic and mathematical symbols and codes of a computer are only a human kind of information processing representing atomic, molecular, cellular and evolutionary processes. According to complexity theory, there are different degrees of computability. This is the hard core of the unified theory of complex networks.

As far as these systems solve problems with a certain degree of complexity, they get degrees of intelligence. The degrees of intelligence in technical systems are traditionally called "artificial intelligence". But, in a technical coevolution, human activity is integrated in sociotechnical systems solving problems with high degrees of complexity. Obviously, in this case, a kind of human-machine intelligence emerges which can no longer be separated into either human or machine intelligence. Consciousness is also a specific state of brains emerging during evolution with certain degrees. But in any case, it is no necessary condition of intelligence. A smart grid, for example, has certain degrees of intelligence and autonomy, but no consciousness. From a scientific point of view, we aim at a unified theory of intelligent systems and human intelligence is only a part of it. But, from an ethical point of view, all these intelligent systems should be initiated and developed as service and assistant systems for human well-being and saving the Earth system. The ethical point of view makes the difference and distinguishes human dignity.

References

acatech (Ed.). (2011). *Cyberphysical systems. acatech position* (acatech = National Academy of Science and Technology). Berlin: Springer.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47–97.

Balch, T., & Parker, L. (Eds.). (2002). *Robot teams: From diversity to polymorphism*. Wellesley: A. K. Peters.

Bekey, G. L. (2005). *Autonomous robots. From biological inspiration to implementation and control*. Cambridge, MA: MIT Press.

Bellman, K. L. (2005). Self-conscious modeling. *IT – Information Technology, 4*, 188–194.

Berners-Lee, T. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by the inventor*. San Francisco: Harper Collins.

Braitenberg, V., & Radermacher, F. J. (Eds.). (2007). *Interdisciplinary approaches to a new understanding of cognition and consciousness*. Universitätsverlag Ulm: Ulm.

Brooks, R. A. (1999). *Cambrian intelligence: The early history of the new AI*. Cambridge, MA: The MIT Press.

Chalmers, D. (2010). *The character of consciousness*. Oxford: Oxford University Press.

CoTeSys. (2006–2011). is funded by the German Research Council DFG as a research cluster of excellence within the "excellence initiative" from 2006–2012.

Cyber-Physical Systems. Program announcements & information. *The National Science Foundation*, Arlington, 30 Sept 2008.

Dreyfus, H. L. (1979). *What computer's can't do – The limits of artificial intelligence*. New York: Harper & Row.

Dreyfus, H. L. (1982). *Husserl, intentionality, and cognitive science*. Cambridge, MA: MIT Press.

Freeman, W. J. (2004). How and why brains create meaning from sensory information. *International Journal of Bifurcation and Chaos, 14*, 515–530.

Friederici, A. D. (2006). The neural basis of language development and its impairment. *Neuron, 52*, 941–952.

Haken, H. (1996). *Principles of brain functioning. A synergetic approach to brain activity, behaviour and cognition*. Berlin: Springer.

Hansmann, U. (2001). *Pervasive computing handbook*. Berlin: Springer.

Hebb, D. O. (1949). *The organization of the behavior*. New York: Wiley.

Lee, E. (2008). Cyber-physical systems: Design challenges. In *University of California, Berkeley Technical Report* No. UCB/EECS-2008-8.

Mainzer, K. (2003). *KI – Künstliche intelligenz. Grundlagen intelligenter systeme*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Mainzer, K. (2007). *Thinking in complexity. The computational dynamics of matter, mind, and mankind* (5th ed.). New York: Springer.

Mainzer, K. (2008a). The emergence of mind and brain: An evolutionary, computational, and philosophical approach. In R. Banerjee & B. K. Chakrabarti (Eds.), *Models of brain and mind. Physical, computational and psychological approaches* (pp. 115–132). Amsterdam: Elsevier.

Mainzer, K. (2008b). Organic computing and complex dynamical systems. Conceptual foundations and interdisciplinary perspectives. In R. P. Würtz (Ed.), *Organic computing* (pp. 105–122). Berlin: Springer.

Mainzer, K. (2009). From embodied mind to embodied robotics: Humanities and system theoretical aspects. *Journal of Physiology, Paris, 103*, 296–304.

Mainzer, K. (2010). *Leben als Maschine? Von der Systembiologie zur Robotik und Künstlichen Intelligenz*. Paderborn: Mentis.

Mainzer, K. (2014). *Die Berechnung der Welt. Von der Weltformel zu Big Data*. München: C.H. Beck.

Mainzer, K., & Chua, L. O. (2011). *The universe as automaton. From simplicity and symmetry to complexity*. Berlin: Springer.

Mainzer, K., & Chua, L. O. (2013). *Local activity principle*. London: Imperial College Press.
Mataric, M., Sukhatme, G., & Ostergaard, E. (2003). Multi-robot task allocation in uncertain environments. *Autonomous Robots, 14*(2–3), 253–261.
Merleau-Ponty, M. (1962). *Phenomenology of perception*. London: Kegan Paul.
Nolfi, S., & Floreano, D. (2001). *Evolutionary robotics. The biology, intelligence, and technology of self-organizing machines*. Cambridge, MA: MIT Press.
Pfeifer, R., & Scheier, C. (2001). *Understanding intelligence*. Cambridge, MA: MIT Press.
Scott, A. (2003). *Nonlinear science. Emergence and dynamics of coherent structures*. Oxford: Oxford University Press.
Shuji Kajita. (2007). *Humanoide roboter. Theorie und Technik des Künstlichen Menschen*. Berlin: Aka GmbH.
Singer, W. (1994). The role of synchrony in neocortical processing and synaptic plasticity. In E. Domany, L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks II*. Berlin: Springer.
Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica, 1*, 261–405.
Thrun, S., Burgard, W., & Fox, D. (2005). *Probabilistic robotics*. Cambridge, MA: MIT Press.
Wedde, H. J., Lehnhoff, S., Rehtanz, C., & Krause, O. (2008). Von eingebetteten Systemen zu Cyber-Physical Systems. Eine neue Forschungsdimension für verteilte eingebettete Realzeitsysteme. In *Pearl 2008 – Informatik Aktuell*. Aktuelle Anwendungen in Technik und Wirtschaft 2007 12.
Weiser, M. (1991). The computer for the 21st century. *Scientific American, 9*, 66–75.
Wilson, E. O. (2000). *Sociobiology: The new synthesis* (25th ed.). Cambridge, MA: Harvard University Press.

Chapter 16
The Anticipatory Brain: Two Approaches

Mark H. Bickhard

Abstract It is becoming increasingly accepted that some form of anticipation is central to the functioning of the brain. But modeling such anticipation has been in several forms concerning what is anticipated, whether and how such 'anticipation' can be normative in the sense of possibly being wrong, the nature of the anticipatory processes and how they are realized in the brain, etc. Here I outline two such approaches – the Predictive Brain approach and the Interactivist approach – and undertake a critical comparison and contrast.

Keywords Predictive brain • Interactivism • Free energy • Anticipatory brain • Brain models

There are at least two approaches to modeling brain functioning that make central use of notions such as anticipation, expectation, and prediction. I will argue, however, that they diverge at very fundamental levels. The first approach that I will address is a broad programme with multiple variants and various names, such as the Predictive Brain, the Bayesian Brain, and the Free-Energy Principle (e.g., Clark 2013; Friston and Stephan 2007). The second is the interactivist model (e.g., Bickhard 2009a, b).

16.1 The Predictive Brain

Predictive brain models have developed within the tradition of Helmholtz's notion of inference from perceptual sensations to representations of the environment (Clark 2013). This sensory empiricist tradition, in turn, extends to the classical Greeks.

A relatively recent major step forward in this tradition was the development of analysis by synthesis models, in which sensory inputs are analyzed via some sort

M.H. Bickhard (✉)
Lehigh University, Bethlehem, PA, USA
e-mail: mhb0@Lehigh.EDU

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_16

of synthesis of those inputs (MacKay 1956, 1969; Neisser 1967). This move in effect substitutes abductive processes for the deductive and inductive processes that Helmholtz depended on.

One important consequence of this development was the introduction of an active brain, one that is ongoingly engaged in the constructive processes of abductive prediction and explanation. This is in contrast to classical reflex arc and other forms of passive, reactive models.

The predictions of an analysis by synthesis process must be in some way compared to the actual pattern and flow of inputs, and this requires that some signals be generated that can be compared in this manner – either successfully or unsuccessfully matching the actual inputs. A natural next elaboration of such models is to recognize that some such internally generated signals might evoke muscle activity, thus action, and thus *induce* the sensory inputs that are being "predicted", not just receive them. This yields models such as *Behavior: The Control of Perception* (Powers 1973). In such models *interaction* becomes central, not just reception.

Computational and information processing models were dominant when analysis by synthesis was introduced, but the general form of predictive synthesis could take other forms as well. One such alternative form that has become widely accepted is that of Bayesian inference. Bayesian inference starts with a prior probability distribution over some space of hypotheses and, taking new data into account, modifies that distribution into a posterior distribution. Such a posterior probability distribution, in turn, can constitute the predictions concerning the (probability distributions) of input patterns and flows. A probability distribution with density concentrated at a point can be taken to predict a single particular input. In cases of real number parameterized spaces of hypotheses, the distributions are characterized in terms of their statistical parameters.[1]

The introduction of Bayesian forms of prediction brings with it the reliance on available prior probability distributions, with important additional model requirements and resources. In standard Bayesian statistics, such priors are often treated as strengths of prior beliefs about the available hypotheses, which are then modified with respect to data. But such prior distributions can also be made dependent on previous data and the distribution manifested in such prior "experience", thus yielding "empirical" prior probability distributions as basis for Bayesian modification into posterior distributions.

One possible source for such prior distributions would be a higher level of Bayesian inference whose posterior distribution parameters could be input as priors to a lower Bayesian process. An iteration of this modeling step yields a hierarchy of Bayesian process layers, with each layer receiving errors of prediction – that which is to be predicted by this layer – *from* below and sending predictions *to* the layer below, with the bottom layer predicting sensory inputs and generating prediction

[1] Sufficient statistics (Friston et al. 2009).

errors relative to those errors to the layer above.[2] Such a hierarchy models multiple layers each of which attempts to predict the inputs from below, generated as errors of prediction at the level of that layer below, thus attempting to account for what lower layers have not been successful in accounting for.

Friston has proposed an integrating framework for such hierarchical predictive models (Friston 2013; Friston and Stephan 2007). In this framework, a statistical notion of "free-energy" substitutes for Bayesian parameter estimates. This is in part a practically motivated substitution, but it also has theoretical implications. Practically, minimal error Bayes estimates are difficult to compute; free-energy is proven to constitute an upper bound for Bayes error, so minimizing free-energy entails (approximately) minimizing Bayes error, and free-energy minimization is a tractable computation (Friston et al. 2012a; Kiebel et al. 2008). Such minimizations are computed, in general, via some sort of descent iterations settling at an available minimum. Minimizing free-energy, thus, accomplishes minimizing prediction error, and thus minimizes "surprise".

Within this framework, Friston proposes that some "predictions" are of proprioceptive inputs, which, in turn, induce muscle activity that yields those inputs, thus inducing action, which in turn induces sensory inputs. This extends the basic hierarchical model to include interactive *inducing* of inputs along with the basic *reception* of inputs.

Within this sensory-action framework, Friston proposes a theoretical modeling use of the reliance on Bayesian prior probability distributions. The hierarchy of Bayesian layers cannot extend unboundedly: there must be some finite number of layers and the highest layer must take its prior distributions from some source other than a still higher layer. It is proposed that they are innate. These highest level prior probability distributions, also called beliefs or expectations, can then be understood as inducing action that yields various kinds of sensory patterns and flows. In this perspective, the induction of action via "expectations" becomes as or more important than simple prediction of input.

In fact, Friston claims that all considerations of properties such as utility, cost, value, and so on can be folded into what counts as surprise relative to these highest level expectations (or beliefs or prior probability distributions – sometimes called hyperpriors). This is because an organism "predicts" what is "expected", and induces those "expected" inputs via action. What is "expected" in innate hyperpriors (so it is claimed) is flows of inputs that are *viable* for the organism, and, thus, that capture or fold-in considerations of utility and value. The argument for this is that, if ancestral organisms did not expect, thus induce, viable flows of action and inputs, they would have died out, thus extant organisms *must* "expect" such viable flows (Kiebel et al. 2008; Friston et al. 2012b).

[2]For an early version of such hierarchical prediction, see Tani and Nolfi (1999).

16.1.1 Some Problems

This is an ambitious, sophisticated, and impressive modeling framework. Nevertheless, I contend that it suffers from significant problems at every level of its structure.

At the broadest level, this programme is a variant of sensory empiricism, based on sensory encodings as inputs. This is a dominant approach to perception, and has been for millennia. Helmholtz is a major figure in this tradition. There are three 'obvious' issues involved in this construal of perception: (1) What is the nature and origin of the inputs upon which the inferences are based? (2) What is the nature of the inference process(es)? And (3) What is the nature and origin of the representations by which the inferred world is characterized?

Classical answers would have it that the inputs are sensations, the inference processes are deduction and induction, and the representations of the wider world are aggregations or organizations of the sensations. In other words, the process is one of inferring world encodings – objects, properties, etc. – on the basis of input encodings. This is basic empiricist epistemology.[3]

Despite millennia of trying, no one has been able to model how causal inputs can generate representational sensory encodings. This is the problem of sensations, sense data, and other basically equivalent terms: the "transduction" of representation from causal inputs (Fodor and Pylyshyn 1981; Bickhard and Richie 1983).[4]

The tradition faces serious problems even if the problem of the mysterious transduction of sensory encodings is overlooked. In particular, not only has no successful model of transduction ever been offered, there has also never been a successful model of how higher level representations could be generated on the basis of such sensory encodings. How, for example, can representations of chair, bachelor, the number three, triangle, or justice be generated or constructed out of sensory encodings?[5]

Bayesian hierarchical models would, in fact, seem to have weaker resources in this regard than classical sensory empiricism. Classical empiricism could attempt to model such representations with various collections and structures of sensory encodings, while the Bayesian hierarchy is limited to parameters of probability distributions over spaces of parameters of probability distributions over … over patterns and flows of inputs.

[3] E.g., Locke, Hume, Russell (in some incarnations), Fodor, much of contemporary literature, and even Aristotle's signet ring impressing its form into wax.

[4] "Transduction, remember, is the function that Descartes assigned to the pineal gland" (Haugeland 1998, p. 223).

[5] Or perhaps they're independently innate? This issue is alive and well in contemporary work: in child development, for example, a fundamental question is whether or not it is possible to construct, say, object encodings or number representations, out of sensory encodings. Some say yes, and some say that such higher-level encoding representations must be innate. Ultimately, neither stance is successful (Allen and Bickhard 2013a, b).

There is, thus, no account of the nature or origins of sensory representations in these models nor of representations of the world in general.[6] Still further, there are in-principle arguments that such accounts are not possible (e.g., Bickhard 2009a). Analysis by synthesis introduces abduction along with deduction, but does not offer anything new regarding an account of representation. Taking into account that action can influence or induce or 'control' perceptual inputs is a major step forward, but, yet again, this offers nothing new regarding the nature of perceptual (or any other) representations.

One aspect of this problem is to note that, just as there is no account of representation, so also is there no account of representational error. There is some sort of comparison of 'predictions' and inputs, and an iterative descent process[7] which tends to minimize overall discrepancies in such comparisons, but such discrepancies constitute error only for an external observer or designer who interprets those inputs as representations and who interprets those discrepancies as error. The machine or organism that iterates in some descent process has no emergent normativity at all, including that of representation and representational error. The sense in which the model might be considered to be a model in which (sensory) representational error is minimized (or any other kind of error) is limited to that of the interpretive perspective of a designer or some other kind of external interpreter.[8]

Yet another aspect of this designer or interpreter dependence of Bayesian hierarchical models is that there is no account of the spaces over which the probability distributions are distributed. These spaces are required and assumed in the nature of Bayesian estimation; their nature and origin is not addressed.

16.1.2 Free-Energy Problems

A central claim for the free-energy model is that it incorporates normative considerations, such as cost or utility, into the higher level expectations – hyperpriors – without having to calculate with respect to such considerations: the organism

[6]This literature proceeds within a background assumption of semantic information models, conflating technical covariation information with representational (about) information, without ever addressing this assumption. It is, nevertheless, evident everywhere, including Clark (2013).

[7]With respect to some underlying metric on the underlying space.

[8]Block and Siegel (2013) suggest that a better term than "error" might be "discrepancy", but this too suggests a normative standard from which the "predictions" are "discrepant". "Difference" is more neutral in this regard: overall, the dynamics of such a system settles into a minimization of such differences. There are no "errors" (Bickhard and Terveen 1995). There are multiple similar abuses of language in this literature, such as "error", "representation", "cause", "belief", "expectation", "describe", etc. none of which (in this literature) refer to anything like the phenomena that such words are generally taken to refer to. Nevertheless, they leave the suggestion, without argument, that they do constitute models for the phenomena at issue (McDermott 1981).

'simply' expects to stay in viable zones of its states, and induces actions that keep it in those zones.

Friston and Stephan (2007) attempt to account for normativity in the model via:

> The basic premise we start with is that biological systems must keep $\tilde{y}$ ["$\tilde{y}$ can be regarded as sensory input"] within bounds (i.e. phase-boundaries) through adaptive changes in α.["effect[s] of the system on the environment"] Put simply, adaptive systems or agents should minimise unlikely or surprising exchanges with the environment. (Friston and Stephan 2007, p. 425)

The idea is that evolutionary selection will eliminate organisms that fail to keep sensory input within bounds, and, thus, to generate organisms that minimize surprise – minimize expectation failure – with respect to those inputs. There is in fact more than one premise in these two sentences, and they are all questionable, but my focus here is on the normativity issue. The purported connection of this model to normativity is the notion of "keeping inputs within bounds". This is not defined, though a supposed characterization is given in "i.e., phase-boundaries". Unfortunately, phase boundaries are part of the infrastructure of all living systems, and changes in phase-boundaries are intrinsic to growth and development. The failure of phase-boundaries to capture what "within bounds" could mean leaves the notion of "not within bounds" hanging as a surrogate for whatever seems not good for the organism from the observer's (or Friston's) perspective. This observer subjectivity of normativity for the organism is not a property of normativity that is fundamental in the constitution of the organism itself (Kiebel et al. 2008).

It is not something intrinsic to the nature of the organism per se, but, rather, is a supposed consequence of the effect of prior evolutionary selection. The organism, in this model, is still 'just' factually and causally settling in to conditions that minimize discrepancies between internally generated signals and input generated signals. There is still no normativity, thus, no error.

16.1.2.1 What Is It to Be a Living Being?

There are allusions to such a model capturing various aspects of self-organization, embodiment, and other properties of living beings, but it offers a strikingly weak characterization of what it is to be alive: tending to stay in a particular set of states.

Friston's model has at best a normativity of avoiding dissolution (Friston 2012) – keeping things "within bounds" (which is not captured by notions of phase-boundaries). Living systems clearly do have to avoid dissolution, but, with this "selection designed" dissolution-avoidance, the model captures less of the nature of living beings than, for example, enactivist notions of living beings being autopoietic – continuously constructing and reconstructing their components (Maturana and Varela 1980; Varela 1979, 1997; Di Paolo 2005; Weber and Varela 2002). This is a real time dynamic characterization, not simply a result of ancestral evolution.

The free-energy grounding claim is in even stronger contrast with the recognition that living beings have to engage with the *environmental thermodynamic and material relationships* for their own continued existence – they have to maintain those relationships in order to self-maintain their own continued existence, in order to recreate components, for example (Bickhard 2009a). Living beings are *constituted* as processes that are *open* systems *of ontological necessity*, not that they *just happen* to be open to inputs from the environment (see Friston's example of a snowflake that is "open" to its environment, Friston and Stephan 2007, p. 423). Living systems are *constituted* as processes that maintain their thermodynamic conditions in sensitivity to changes in environmental conditions (Bickhard 2009a). They *could not* be closed systems, whereas a snowflake, in principle, could. The normativity of living systems, including of their representing, emerges in this necessity to maintain, create, obtain, and exploit the conditions and the material for their own existence (Bickhard 2009a), not just to avoid dissolution.[9]

This property of the self-maintenance of the conditions for ones own continued existence is a much deeper condition than simply "avoiding dissolution because ancestors were selected for doing so". Selection histories do not suffice: for example, some crystals grow in a certain way because earlier deposition of molecules from solution happened to realize a particular form of crystal structure, and the seed crystal thus created selects further depositions to fit that form – this is a case of historistic selection (Bickhard and Campbell 2003), but there is no life and no normativity: neither a molecule nor the crystal are alive, nor is either *in error* if the molecule should happen to deposit in a position or orientation consistent with a different crystalline form. (Evolutionary) selection histories can *create* systems that are alive and with emergent normativity, but neither being alive nor that normativity is *constituted* in having such a selection history.

16.1.2.2 Error Minimization

Overlooking problems concerning the nature of normativity, there are seemingly obvious problems for an error minimization stance, such as the "dark room problem" (Clark 2013): If the organism seeks to minimize prediction error with regard to inputs, why doesn't it just head to a dark room and stay there? Input prediction error would seem to be fully minimized when there are no inputs. There is a seemingly ready answer within the hyperprior framework: the organism *expects* to be in lighted areas, and so turns on the light in order to fulfill that expectation (Friston et al. 2012a; Friston 2013). This is a straightforward extension of the basic notion that all normative considerations are built in to the hyperpriors (Friston et al. 2012b).

[9]This connection between cognition and life was at the center of the interactivist model from its inception, e.g., "knowing as explicated above is an intrinsic characteristic of any living system" (Bickhard 1973, p. 8; also Bickhard 1980a, p. 68). This is also a strong intuition in the enactivist framework, but, so I argue, is not so well captured by the definition of autopoiesis.

But there will be exceptions, such as if the organism wants darkness, perhaps in order to sleep or hunt or hide. These exceptions too can be built-in to the hyperpriors, but a problem begins to be discerned: how many exceptions and exceptions to the exceptions have to be built-in in order to accommodate all normative considerations into hyperpriors? How could they all possibly be 'built-in'? Isn't such an epicyclic elaboration of hyperpriors a *reductio* of the approach?

Another example would be an animal on a shock grid. When the tone that signals impending shock is heard, why doesn't the animal 'simply' predict pain and stay on the grid? Why would it learn to jump off of the grid? Perhaps absence of pain is built into the hyperpriors? But what about the exceptions of seeking pain inputs, such as with hot peppers? Are those also 'built-in' to the hyperpriors?

Hyperprior 'expectations' seem to handle normative phenomena only insofar as all relevant normative considerations are already built-in to those hyperpriors. That does not constitute an adequate explanation or model (Gershman and Daw 2012; Roesch et al. 2012). Thus, there is no model of how or why an organism avoids harm or seeks value. It is simply assumed that it does so as a result of training or evolution.[10] For organisms, then, the assumption is that an adequate set of expectations to able to account for all normative behavior and behavior learning is somehow evolved into the hyperpriors.

16.1.2.3 "Built-in" to What?

Friston at times writes as if these highest level expectations are somehow 'built-in' to the organization of the whole organism, not just the brain (e.g., Friston et al. 2009). Such a perspective is clearly more powerful than the assumption that they are innate in the brain, but the most obvious manner in which this could be the case is for the organism to be 'organized' so that it tends to avoid hunger and pain etc., and tends to seek pleasure and excitement, etc. But such explicit cost and utility considerations and their underlying processes and neural realizations are precisely what is claimed to be obviated by the hyperprior model.

These claims and the mathematics that support them are mostly elaborations of mathematical equivalences to the basic assumption that the organism tends to stay in some set of 'expected' states, and that those states will tend to be viable states whose expectations are induced by evolution (Friston 2012). That basic set of 'expected' states, then, has to accommodate all normative considerations.

[10]But training has to be with respect to some normative criteria, and there are none other than what is built-in to the hyperpriors.

16.1.2.4 Learning

Hierarchical Bayesian prediction models have powerful resources for learning: hierarchies of Bayesian layers, each generating predictions of, and thus accommodating, the errors of prediction from layers below. Each layer can learn *sequences* of the sufficient statistic parameters for probability distributions over the spaces of the layers below, which, in turn, can learn sequences of the layers below them. So the Bayesian layers can collectively learn sequences of sequences of ... of patterns and flows of inputs (Kiebel et al. 2008) (perhaps as induced by action 'predictions').

The highest level of such expectations, hyperpriors, is at times characterized as a set-point (e.g., Friston et al. 2012b), and the overall hierarchy does have a flavor of a servomechanism hierarchy, in which higher levels send (sequences) of goals to lower levels. But learning sequences of sequences cannot capture the unbounded spaces of possible interactions that even a simple feedback servomechanism can manifest. The Bayesian hierarchy can generalize beyond actually experienced flows of interaction, but only in terms of sequences of sequences related by the descent processes used to calculate minimum free-energy solutions.

It is crucial to note, in this regard, that the underlying spaces upon which these calculations take place, as well as the metric organizations on those spaces that are necessary for any generalizations to take place – for descent calculations to proceed – must be already available for the calculations to take place. That is, they must be innate. The Bayesian processes offer no way in which new such spaces or new such metrics can be generated, so the spaces and their metrics must not only be innate, they are fixed (e.g., Kiebel et al. 2008): new layers cannot be generated and new metrics, or any other kind of topological or metric organization, cannot be organized. New layers of generalization, thus, as well as the metrics and topologies of generalization – and thus new kinds of generalization – cannot be learned. This holds at all layers of a Bayesian hierarchy, not just the highest hyperprior layer. Yet we know that analogy and metaphor, for example, can induce new cognitive organizations, can re-organize similarity spaces – that is, can change *via learning* the topological and metric organizations involved (Gentner and Rattermann 1991; Gentner and Jeziorski 1993; Medin et al. 1993).

It would be possible for one layer to switch a lower layer from one space to another, perhaps with a different metric (Friston et al. 2009), but, again, only if the spaces to be switched among are already innately available in the lower layer. Again, this is an example in which everything has to be already pre-prepared for whatever will be needed.

16.1.2.5 CNS Architecture

Friston claims that the Bayesian layers of a hierarchical Bayesian prediction process are realized in brain architecture, and, thus, that hierarchical Bayesian prediction layers constitute the basic functional architecture of the brain (Friston and Stephan 2007; Kiebel et al. 2008; Adams et al. 2012). Such a characterization

of CNS architecture makes partial sense for some cortical domains in early sensory processing, especially for the visual system, though it is not (yet) clear that the microfunctioning of those 'layers' fits the model (Clark 2013; Friston 2008), and it is argued that motor functioning can also be (partially) accommodated (Adams et al. 2012).

With respect to larger and more general considerations, however, the hierarchical Bayesian model does not fit well. For example, there are large parts and regions of the brain that do not have a hierarchical organization, such as multiple node loops involving cortex and sub-cortical regions (Doya 1999; Koziol and Budding 2009; Nieuwenhuys 2001)[11] – such as prefrontal to striatum to thalamus to prefrontal, or prefrontal to cerebellum to thalamus to prefrontal, and so on. Multiple-node loops abound and are not consonant with a simple hierarchy.

Furthermore, there are multifarious characteristics of brain functioning that are not touched upon by the neural network hierarchical Bayesian model. These include silent neurons that rarely or never fire, gap junctions, intrinsic oscillatory activity (not just iterative re-entrance) at both neural and circuit levels, wide-spread release of neuro-modulators that can functionally reorganize activity (Doya 2002; Marder and Thirumalai 2002; Marder 2012), emotional processes and the limbic portions of the brain that underlie them, episodic and biographical memory and learning, and so on. Most of what we know about the brain is either not accounted for by these models or is flatly inconsistent with them.

16.1.2.6 So, Hierarchical Bayes?

Hierarchical Bayes, thus, is simply not powerful enough to account for multiple phenomena of normative activity, of learning and development, of neural functioning at both micro- and macro-levels, and how any of these are related to the nature of being alive.

There are fundamental problems resident in the basic framework of sensory encoding assumptions about the nature of representation; problems concerning the nature of representation are not addressed in these models. Additions to the sensory inference model in the form of abductive anticipation, action as inducing sensory inputs, hierarchical Bayesian processes of prediction, and free-energy claims to account for normative phenomena and for brain functional architecture each add additional power to the resources of the overall model, but do not resolve any of the difficulties inherited from earlier model innovations, and each introduces new problems of its own.

I will argue, nevertheless, that there are crucial insights in these models that need to be maintained – in particular, the importance of anticipation and of *inter*action, not just inputs and action – and I outline a model that arguably does so without encountering the problems of free-energy hierarchical Bayes.

[11] In spite of brief mention of such architectures in Adams et al. (2012).

16.2 The Interactivist Model

Predictive encoding models focus on prediction of inputs at the sensory interface. In contrast, consider the possibility that it is *flow of interaction* that is anticipated, not the inputs (or outputs) that participate in that interaction. A perspective on this possibility can be found at the root of the interactivist model:

> Consider two Moore machines [abstract finite state machines with outputs] arranged so that the outputs of each one serve as the inputs of the other. Consider one of the Moore machines as a system and the other as its environment, and let the system have the initial and final state selections that make it a recognizer.
>
> The system can thus recognize input strings in the standard sense in automata theory [a recognizer "recognizes" strings of inputs that move it from its initial state to one of its final states]. In this interactive configuration, however, an input string corresponds to – is generated by – a state transition sequence in the environment. The set of recognizable input strings thus corresponds to the particular set of state sequences in the environment that could generate them. The recognition, or knowing, relationship is thus extended from inputs to situations and conditions in the environment.
>
> Furthermore, during an interaction, the environment is receiving outputs from the system – and it is these outputs from the system that induce the environmental state transitions that generate the inputs to the system that the system either recognizes or doesn't. Thus the 'recognition' process is no longer strictly passive – the 'recognized' strings are induced from the environment by the system's own outputs. In fact, the interaction doesn't need to be viewed as a recognition process at all. It is equally as much a construction or transformation process – constructing the situations and conditions corresponding to the last state of a 'recognizable' environmental state sequence – or at least a detection process – detecting an initial state of a 'recognizable' environmental state sequence – and so on.
>
> The system need not be thought of as a single undifferentiated recognizer. It could be, for example, a collection of recognizers connected to each other, say, with the final states of one attached to the initial state of another. Such connections could induce functional relationships among the recognizers, such as one testing for the appropriate conditions for another to begin, or a servomechanism being used to create a subcondition for another process to proceed, etc. (Bickhard 1973, pp. 21–22: also in Bickhard 1980a, pp. 75–76)

Here we have a formalization within abstract machine theory of the idea of recognizing input strings, and of inducing those "recognizable" input strings from the environment. This sounds a bit like the prediction and control of inputs of action oriented predictive processing (Clark 2013). And it is.

But the contrasts are crucial. First, the inputs in the interactivist model are not mis-construed as representational; they are not supposed to be transductions-into-representations. Instead, they are simply registrations that move the machine in its state transition diagram. Representation emerges in the implicit detection or presupposition relationships between the induction of recognizable strings and the environmental conditions that support such recognizable string induction. This is a fundamentally different conception of representing, and one that does not encounter the myriad problems of encodingisms. ***It is the anticipating that is***

representational – truth valued – ***not the inputs***. This is a pragmatic, action based model of representation, and it is much stronger than any encodingist model.[12]

Note in particular that if there is a connection from one recognizer to another, then the 'anticipation' that some environment recognized or transformed by the first will also be one that is 'recognizable' by the second could be false, and could be discovered to be false in virtue of the second recognizer not entering one of its final states (Bickhard 1980a), perhaps, for example, encountering a halting condition instead. Another consequence is that, because inputs are not representational, the representation relationship is not restricted to things that can be 'represented' by aggregations and organizations of inputs.

The second difference is that the abstract machine configuration is not in itself normative. There is no normative difference from the machine perspective between 'successfully' recognizing or inducing relevant input strings or 'failing' to do so. All such processes and outcomes are 'just' factual, 'causal', processes in the machine(s) and environment. Accounting for normativity requires further fundamentally thermodynamic considerations (Bickhard 1993, 2009a) that are not present in the abstract machine model. Predictive encoding models, in contrast, treat the inputs as transductions and expectations as normative, when in fact they too involve nothing more than factual, causal processes. With no normativity, there is no error; predictive encoding models settle into conditions in which they minimize discrepancies between inputs and internally generated signals that are matched against those inputs. They are a complicated form of unsupervised "learning" (Bickhard and Terveen 1995), but do not transcend the basic limitations of such models.

16.2.1 What and Where of Anticipation

Predictive brain models focus on anticipating or inducing global inputs, while the interactivist model focuses on anticipating interactive flow. Anticipating interactive flow has local aspects, as well as global aspects. In particular, local domains of the brain have to successfully anticipate their own local flow of process in order for the whole brain to successfully anticipate its global, thus interactive, processes. In shifting from anticipation of inputs to anticipation of interactive processing, thus, we also shift from global perspectives to local perspectives.

In particular, each local domain of the brain engages in anticipatory processes concerning its own local near future processes. This is a primary focus of the interactivist model. Local anticipatory processes are realized in local "set-ups" of preparation for particular ranges of potential further process flow – a "microgenesis"

[12]For discussions of action based anticipatory models, see, for example, Bickhard (1980a, b, 1993, 2009a, b), Bickhard and Richie (1983), Bickhard and Terveen (1995), Buisson (2004), Pezzulo (2008), Pezzulo et al. (2013).

of preparation of local conditions for further processes (Bickhard 2009c, 2015a, b, in preparation). Such microgenesis, thus, anticipates that the future will remain consistent with what has been microgenetically set-up for.

Of central importance is that microgenetic anticipation can be correct or not correct: true or false. This is the locus of the emergence of representational truth value. Much more needs to be done to model more complex forms of representation, but truth value is the central barrier to naturalistic models of representation (for further development of the representational model, see, e.g., Bickhard 1980b, 1993, 2009a, 2015a, b, in preparation).

16.2.2 Learning

Learning is a major integrating perspective in this model for brain functioning and brain evolution (Bickhard 2015a, b). Here I will introduce one functional assumption and argue that it suffices to account for multiple forms of learning.

That assumption is that successful microgenetic anticipation yields consolidation of those microgenetic processes in those functional circumstances, while unsuccessful microgenetic anticipation yields destabilization. This models a variation and selection process – an evolutionary epistemology – with successful (microgenetic) anticipation as the selection criterion (Campbell 1974; Bickhard and Campbell 2003).

16.2.2.1 Habituation

If some part of the nervous system has a simple matching or subtractive relationship with its inputs, then microgenetic anticipation of the processes of input registration will amount to anticipation of those inputs *per se*. This is the basic model for Sokolov habituation (Sokolov 1960). For a simple tone, this requires only the first cochlear nucleus to be able to successfully anticipate.

16.2.2.2 Classical Conditioning

If pain inputs are inputs that cannot be habituated, and, furthermore, offer no possibility of successful interaction (to a first approximation), then the only way to successfully interact with a sequence of tone followed by (say) shock is to avoid the shock – jump off of the shock grid. Note that there are no “hyperprior” “expectations” involved.

16.2.2.3 Instrumental Conditioning

If the hypothalamus is generating signals as a result of low blood sugar, then the only way to (ultimately) interact with this input is to do something that raises blood sugar.[13] These activities can be enormously complex, as well as contextually and culturally variable, though fussing and crying generally suffice for infants (together with supportive forms of interaction such as rooting and sucking, as well as a supportive environment). In any case, learning will stabilize on forms of interaction that successfully halt or diminish the hypothalamic signals.

16.2.2.4 Other Forms of Learning

These forms of learning all involve successful termination or diversion of an input stream. But this is not the only manner in which microgenetic anticipation can be successful. If microgenesis can proceed in a temporal trajectory that ongoingly successfully anticipates the ongoing trajectory of process flow, then that too manifests successful microgenetic anticipation. These forms of microgenetic anticipation underlie forms of learning such as incidental learning, memory, and so on (Bickhard 2006, 2015a, b). I will not elaborate these points here,[14] but will proceed to some further comparisons with predictive brain models.

16.2.3 Partial Convergences

Predictive brain models can account for Sokolov habituation very directly: actual inputs are matched by predicted inputs, and the predictions are successful. If the only relevant activity of some part of the brain is to register inputs, then local microgenetic anticipation is extensionally equivalent to input (registration) prediction, and there is a close convergence between interactivist and predictive brain models. They are both models of anticipatory or predictive processes at the center of brain functioning, and the two different senses of anticipation/prediction have a strong convergence for habituation.

The difference between predicting inputs and anticipating local microgenetic processes is subtle in this case, but yields wide divergences for more complex cases. In effect, predictive brain models assimilate *all* brain processes to complex habituation (successful input subtraction) processes.

[13] Hunger and eating is much more complex than this, with multiple feedforward and feedback processes, but this captures the basic organization of the phenomenon (Carlson 2013).

[14] This paper is not the occasion to attempt to present the entire model. I have addressed only enough to be able to make some comparisons with predictive brain models.

Classical conditioning begins to manifest a wider divergence. For the interactivist model, jumping off of the shock grid is a successful way to interact with, thus successfully anticipating the interactive process flow, a tone-to-shock input stream: prevent the shock. The free-energy version of the predictive brain models must hypothesize some sort of innate hyperprior "expectation" that the organism will not experience shock or pain, which then yields some sort of steepest descent convergence on a behavior of jumping off the grid.[15] This, as mentioned earlier, is ad-hoc and has difficulty accounting for exceptions such as for eating hot peppers. The interactivist model has a ready resource for accounting for such exceptions: if the organism has learned to interact with some pain inputs, such as from peppers, with ongoing microgenetic anticipation of the processes evoked by those inputs, then that too is successful anticipatory interaction.[16,17]

To reiterate: the focus of predictive brain models is on the global organism interface with sensory inputs, extended to include proprioceptive "inputs" as a way of accommodating action. In these models, representation is somehow based on the inputs as sensory representations, as well as on the innate spaces, metrics, and probability distributions over those spaces as layers of a Bayesian hierarchy are taken into consideration. In the interactive model, *it is the anticipating that is representational*, not the inputs. The inputs, as well as outputs, play an essential role in influencing the processes that microgenesis is "attempting" to anticipate, but neither inputs nor outputs per se are representational in this model. Among other consequences is that the model completely transcends the problem of accounting for some sort of transductive interface between organism and world.

[15] Note that "steepest descent" processes are not nearly as general as an evolutionary epistemology. This is another aspect of the fact that the Bayesian models require pre-given spaces, metrics on those spaces, and innate distributions (expectations) at least at the highest "hyperprior" level.

[16] And such forms of interaction – e.g., with peppers – will not evoke negative emotional processes, such as fear and anxiety. I will not present the interactivist model of emotions here, but wish to point out that they too are involved in successful microgenetic anticipation (Bickhard 2000).

[17] Insofar as the highest level hyperpriors are "built-in" to the whole organism, not just into the nervous system, it might be claimed that such properties of pain inputs are what constitute the relevant hyperprior(s) for pain. But the interactivist model for pain and for learning with respect to pain is a selection model, a cost or utility or normative model, which – as mentioned earlier – is precisely what Bayesian hyperpriors are supposed to obviate the need for. Thus, to make such a claim contradicts the supposed ability of the Bayesian hierarchical predictive brain model to do without explicit cost or norm considerations.

16.2.4 Functional Processes in the Brain

At this point, I will present a brief overview of how interactive processes are realized in the brain.[18] As will be seen, this view accommodates multiple phenomena that are known about brain functioning, but are at best anomalous for standard views. The overview will be in two parts: a focus on micro-functioning and a focus on macro-functioning.

16.2.4.1 Micro-functioning

The interactivist model focuses on interaction as the fundamental locus for the emergence of representation and cognition (Bickhard 2009a).[19] This ranges from interactions of organisms with their environments to interactions of local brain regions with their surrounds, including other parts of the brain, the body, and the environment.

One consequence of a focus on interaction is the necessity of accounting for timing: successful interactions must having the right timing relationships with whatever is being interacted with in order for those interactions to be successful.[20] Timing, in turn, requires clocks, but clocks are "just" oscillators and the easy way to accommodate the need for oscillators is for all functional processes to be realized as oscillatory processes modulating each other. This modeling framework is *at least as powerful* as Turing machines: a limit case of modulation is for one process to switch another on or off, and Turing machines can be constructed out of switches. It is *more powerful* than Turing machine theory in that it has inherent timing.

If we examine brain processes, multifarious forms of endogenously active oscillatory and modulatory relationships is precisely what we find. These range from the spatially small and temporally fast, such as gap junctions, to large and slower, such as volume transmitters and astrocyte influences:

- silent neurons that rarely or never fire, but that do carry slow potential waves (Bullock 1981; Fuxe and Agnati 1991; Haag and Borst 1998; Roberts and Bush 1981);
- volume transmitters, released into intercellular regions and diffused throughout populations of neurons rather than being constrained to a synaptic cleft (Agnati et al. 1992, 2000); such neuromodulators can reconfigure the functional prop-

[18] See, for example, Bickhard (1997, 2015a, b, in preparation; Bickhard and Campbell 1996; Bickhard and Terveen 1995).

[19] In thus focusing on action and interaction, the interactivist model is in strong convergence with Piaget and with other pragmatist influenced models (Bickhard 2006, 2009a). (There is in fact an intellectual descent from Peirce and James through Baldwin to Piaget.) There are also interesting convergences of this model with Dewey.

[20] Timing goes beyond sequence, and, thus, goes beyond Turing machine theory and equivalents (Bickhard and Richie 1983; Bickhard and Terveen 1995).

erties of "circuits" and even reconfigure functional connectivity (Marder and Thirumalai 2002; Marder 2012);
- gaseous transmitter substances, such as NO, that diffuse without constraint from synapses and cell walls (e.g., Brann et al. 1997);
- gap junctions, that function extremely fast and without any transmitter substance (Dowling 1992; Hall 1992; Nauta and Feirtag 1986);
- neurons, and neural circuits, that have resonance frequencies, and, thus, can selectively respond to modulatory influences with the "right" carrier frequencies (Izhikevich 2001, 2002, 2007);
- astrocytes that[21]:
  - have neurotransmitter receptors,
  - secrete neurotransmitters,
  - modulate synaptogenesis,
  - modulate synapses with respect to the degree to which they function as volume transmission synapses,
  - create enclosed "bubbles" within which they control the local environment within which neurons interact with each other,
  - carry calcium waves across populations of astrocytes via gap junctions.[22]

Thus we find a vast spatial and temporal range of oscillatory and modulatory processes, *all* of which are anomalous from any kind of passive threshold switch or connectionist node modeling perspective.[23]

Crucial to my current purposes is that the spatially larger and (thus) temporally slower processes influence the smaller faster processes – which include the classical synaptic influences – by modulating the local environments, such as ion and transmitter concentrations.[24,25] The slower time scales of such processes imply that they, in effect, set parameters for the faster processes. At the faster scales, these parameters are approximately constant, though they undergo their own trajectories of change at those slower scales. Parameter setting for endogenously active dynamic

[21] The literature on astrocytes has expanded dramatically in recent years: e.g., Bushong et al. 2004; Chvátal and Syková 2000; Hertz and Zielker 2004; Nedergaard et al. 2003; Newman 2003; Perea and Araque 2007; Ransom et al. 2003; Slezak and Pfreiger 2003; Verkhratsky and Butt 2007; Viggiano et al. 2000.

[22] This list from Bickhard (2015a, b). It is *not* an exhaustive list of the multifarious forms of functioning in the brain.

[23] For a model that also addresses some of these scale phenomena, see, e.g., Freeman (2005; Freeman et al. 2012).

[24] For a discussion of relatively local volume transmitter influences, often characterized as neuromodulators, see Marder (2012; Marder and Thirumalai 2002). This is 'just' one class of larger scale, slower forms of modulation.

[25] Kiebel et al. (2008) discuss differential times scales involved in the Bayesian hierarchical model, but, in that model, the time scale differences arise because differing sequences being tracked in the environment change on differing time scales, not because of any differences at the neural and glial level (Bickhard 2015a, b).

processes is the dynamic equivalent of programming for discrete, one step at a time, computational models. The larger, slower, processes, thus, "program" the faster processes: they "set them up" for what those faster processes will be doing in the near future. They constitute microgenesis processes, and, thus, microgenetic anticipation processes.

Note that, just examining what we know about brain functioning, we find these larger, slower processes that set parameters for smaller, faster processes. That is, we find microgenetic "programming". This is intrinsically anticipative, and, thus, intrinsically has truth value – and we find that the interactive model of representation entails brain properties that we in fact find, and, in reverse, the range of brain processes entail the kind of anticipative processes that constitute the interactive model of representation. This reciprocal entailment is a strong consilience.

Local processes, then, especially in the cortex (Bickhard 2015a), realize the sort of microgenetic anticipatory processes that constitute emergent representation. What modulates those local processes?

16.2.4.2 Macro-functioning

Local processes are modulated by more global processes. These too will be oscillatory/modulatory processes engaged in reciprocal projections among cortical regions, reciprocal projections between thalamus and cortex, multi-node loops involving subcortical regions, such as prefrontal to striatum to thalamus to prefrontal, and so on. Modulation relationships among oscillatory processes are not only inherent in single cell and local processes, but also in the general two and more node loops that make up macro-brain architecture (Koziol and Budding 2009). Differing loops in this architecture will manifest differing sorts of influence on the rest of the brain, such as organization of interaction, including sensory interaction, apperception of the situation in which the organism is located, conditions of dynamic uncertainty concerning what to do next, and so on.[26]

One crucial issue is how the brain arrives at any sort of functional coherence in its activities. How does it achieve functional coherence in the manner in which it engages in its multiple macro-level interactions, and in which it modulates myriad local microgenesis processes? One seemingly obvious answer might be that some central executive controls what the rest of the brain is to do, likely the prefrontal

[26]For more specificity concerning such macro-functional considerations, see Bickhard (2015a, b, in preparation). For the general model of perception, apperception, and so on, see Bickhard and Richie (1983; Bickhard 2009a). The model of perceiving offered has strong convergences with Gibson (1966, 1977, 1979), but also some important divergences (Bickhard and Richie 1983). A partial convergence with the model of interactive knowledge of the situation is found in Gross et al. (1999).

cortex. But this "answer" begins a regress of executive decision making: how does the prefrontal cortex achieve its own functional coherence – how does it decide what to do?[27]

Each process in the brain will tend to recruit other processes into modes that yield overall successful interaction and anticipation of successful interaction. In that sense, each process is competing with other processes to generate a self-organization of brain activity into a functioning form that responds to overall interactive success. Brain regions that are not recruited in such self-organizing activity will simultaneously be induced to engage in learning that stabilizes when the region *can* participate in such self-organization, and it will tend to induce such learning changes in other brain processes so that they will tend to generate successful processes in accordance with that local region – that is, the general nature of the processes of competitive recruitment that yields self-organization.

The prefrontal cortex is ideally suited to facilitate such self-organizing integrative processes, but does not have to function as a supreme executive to do so. Instead, it is a locus at which many functional loops converge in such a way that they can participate in global self-organization.[28]

So, multiple domains of the brain are active in differentiating differing aspects of the internal and external situation. Each competes to recruit other domains to interact with its dynamics, thus, with whatever it is differentiating. Striatum loops, cerebellum loops, limbic loops, etc. all specialized for recruiting for special aspects of the situation (Koziol and Budding 2009) – thus (when successful)[29] inducing a kind of self-organization of macro-functioning. This is in strong contrast to the Bayesian brain model of a fixed hierarchy of layers.[30]

16.3 Conclusion

Interactive anticipation is the central nature of the emergence of representation. But it is the anticipating that is representational – truth valued – not the inputs (or outputs) that influence internal processes. Global sensory empiricism is the wrong framework for modeling brain anticipatory processes. The brain overall, and each

[27] Certainly not via some set of fixed innate hyperpriors.

[28] There is, of course, no guarantee that such self-organization will (fully) succeed at any particular time or in any particular situation.

[29] Lack of coherence is certainly possible, and it can also be functional (in several ways) for the brain to engage in chaotic processes. For example, chaotic processes can be a baseline form of process from which functional attractor landscapes can be induced and controlled (Freeman 1995, 2000a, b; Freeman and Barrie 1994; Bickhard 2008).

[30] Note also that in the Bayesian brain model, the reciprocal projections among various cortical regions are supposed to be engaging in descent iterations, not oscillations (Friston et al. 2012b).

local domain of the brain, engage in "attempting" to anticipate future process, via interaction and microgenesis, and microgenetic recruitment of other domains.

The interactive model and various predictive brain models converge on the centrality of anticipation, but strongly diverge in what that means and how it is manifest in brain functional activity.

References

Adams, R. A., Shipp, S., & Friston, K. J. (2012). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*. doi:10.1007/s00429-012-0475-5.

Agnati, L. F., Bjelke, B., & Fuxe, K. (1992). Volume transmission in the brain. *American Scientist, 80*(4), 362–373.

Agnati, L. F., Fuxe, K., Nicholson, C., & Syková, E. (2000). *Volume transmission revisited* (Progress in brain research, Vol. 125). Amsterdam: Elsevier.

Allen, J. W. P., & Bickhard, M. H. (2013a). Stepping off the pendulum: Why only an action-based approach can transcend the nativist-empiricist debate. *Cognitive Development, 28*, 96–133.

Allen, J. W. P., & Bickhard, M. H. (2013b). The pendulum still swings. *Cognitive Development, 28*, 164–174.

Bickhard, M. H. (1973). *A model of developmental and psychological processes*. Ph.D. Dissertation, University of Chicago.

Bickhard, M. H. (1980a). A model of developmental and psychological processes. *Genetic Psychology Monographs, 102*, 61–116.

Bickhard, M. H. (1980b). *Cognition, convention, and communication*. New York: Praeger Publishers.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental & Theoretical Artificial Intelligence, 5*, 285–333.

Bickhard, M. H. (1997). Cognitive representation in the brain. In R. Dulbecco (Ed.), *Encyclopedia of human biology* (2nd ed., pp. 865–876). San Diego: Academic.

Bickhard, M. H. (2000). Motivation and emotion: An interactive process model. In R. D. Ellis & N. Newton (Eds.), *The caldron of consciousness* (pp. 161–178). Amsterdam/Philadelphia: J. Benjamins.

Bickhard, M. H. (2006). Developmental normativity and normative development. In L. Smith & J. Voneche (Eds.), *Norms in human development* (pp. 57–76). Cambridge: Cambridge University Press.

Bickhard, M. H. (2008, May 22–23). The microgenetic dynamics of cortical attractor landscapes. *Workshop on Dynamics in and of Attractor Landscapes*, Parmenides Foundation, Isola d'Elba, Italy.

Bickhard, M. H. (2009a). The interactivist model. *Synthese, 166*(3), 547–591.

Bickhard, M. H. (2009b). Interactivism. In J. Symons & P. Calvo (Eds.), *The routledge companion to philosophy of psychology* (pp. 346–359). London: Routledge.

Bickhard, M. H. (2009c). The biological foundations of cognitive science. *New Ideas in Psychology, 27*, 75–84.

Bickhard, M. H. (2015a). Toward a model of functional brain processes I: Central nervous system functional micro-architecture. *Axiomathes*. doi:10.1007/s10516-015-9275-x.

Bickhard, M. H. (2015b). Toward a model of functional brain processes II: Central nervous system functional macro-architecture. *Axiomathes*. doi:10.1007/s10516-015-9276-9.

Bickhard, M. H. (in preparation). *The whole person: Toward a naturalism of persons – Contributions to an ontological psychology*.

Bickhard, M. H., & Campbell, R. L. (1996). Topologies of learning and development. *New Ideas in Psychology, 14*(2), 111–156.

Bickhard, M. H., & Campbell, D. T. (2003). Variations in variation and selection: The ubiquity of the variation-and-selective retention ratchet in emergent organizational complexity. *Foundations of Science, 8*(3), 215–282.
Bickhard, M. H., & Richie, D. M. (1983). *On the nature of representation: A case study of James Gibson's theory of perception*. New York: Praeger Publishers.
Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Amsterdam: Elsevier Scientific.
Block, N., & Siegel, S. (2013). Attention and perceptual adaptation. *Behavioral and Brain Sciences, 36*, 205–206.
Brann, D. W., Ganapathy, K. B., Lamar, C. A., & Mahesh, V. B. (1997). Gaseous transmitters and neuroendocrine regulation. *Neuroendocrinology, 65*, 385–395.
Buisson, J.-C. (2004). A rhythm recognition computer program to advocate interactivist perception. *Cognitive Science, 28*(1), 75–87.
Bullock, T. H. (1981). Spikeless neurones: Where do we go from here? In A. Roberts & B. M. H. Bush (Eds.), *Neurones without impulses* (pp. 269–284). Cambridge: Cambridge University Press.
Bushong, E. A., Martone, M. E., & Ellisman, M. H. (2004). Maturation of astrocyte morphology and the establishment of astrocyte domains during postnatal hippocampal development. *International Journal of Developmental Neuroscience, 2*(2), 73–86.
Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 413–463). LaSalle: Open Court.
Carlson, N. R. (2013). *Physiology of behavior* (11th ed.). Upper Saddle River: Pearson.
Chvátal, A., & Syková, E. (2000). Glial influence on neuronal signaling. In L. F. Agnati, K. Fuxe, C. Nicholson, & E. Syková (Eds.), *Volume transmission revisited* (Progress in brain research, Vol. 125, pp. 199–216). Amsterdam: Elsevier.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253.
Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429–452.
Dowling, J. E. (1992). *Neurons and networks*. Cambridge, MA: Harvard University Press.
Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? *Neural Networks, 12*, 961–974.
Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks, 15*, 495–506.
Fodor, J. A., & Pylyshyn, Z. (1981). How direct is visual perception?: Some reflections on Gibson's ecological approach. *Cognition, 9*, 139–196.
Freeman, W. J. (1995). *Societies of brains*. Mahwah: Erlbaum.
Freeman, W. J. (2000a). *How brains make up their minds*. New York: Columbia.
Freeman, W. J. (2000b). *Mesoscopic brain dynamics*. London: Springer.
Freeman, W. J. (2005). NDN, volume transmission, and self-organization in brain dynamics. *Journal of Integrative Neuroscience, 4*(4), 407–421.
Freeman, W. J., & Barrie, J. M. (1994). Chaotic oscillations and the genesis of meaning in cerebral cortex. In G. Buzsaki, R. Llinas, W. Singer, A. Berthoz, & Y. Christen (Eds.), *Temporal coding in the brain* (pp. 13–37). Berlin: Springer.
Freeman, W. J., Livi, R., Obinata, M., & Vitiello, G. (2012). Cortical phase transitions, non-equilibrium thermodynamics and the time-dependent Ginzburg-Landau equation. *International Journal of Modern Physics B, 26*(6), 29 p.
Friston, K. J. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211. doi:10.1371/journal.pcbi.1000211.
Friston, K. J. (2012). A free energy principle for biological systems. *Entropy, 14*, 2100–2121. doi:10.3390/e14112100.
Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences, 36*, J.212–J.213.
Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159*, 417–458.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS ONE, 4*(7), e6421. doi:10.1371/journal.pone.0006421.

Friston, K. J., Adams, R. A., Perrinet, L., & Breakspear, M. (2012a). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 1–20.

Friston, K. J., Samothrakis, S., & Montague, R. (2012b). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*. doi:10.1007/s00422-012-0512-8.

Fuxe, K., & Agnati, L. F. (1991). Two principal modes of electrochemical communication in the brain: Volume versus wiring transmission. In K. Fuxe & L. F. Agnati (Eds.), *Volume transmission in the brain: Novel mechanisms for neural transmission* (pp. 1–9). New York: Raven.

Gentner, D., & Jeziorski, M. (1993). The shift from metaphor to analogy in western science. In A. Ortony (Ed.), *Metaphor and thought* (2nd ed., pp. 447–480). New York: Cambridge University Press.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 225–277). London: Cambridge University Press.

Gershman, S. J., & Daw, N. D. (2012). Perception, action, and utility: The tangled skein. In M. I. Rabinovich, K. J. Friston, & P. Verona (Eds.), *Principles of brain dynamics: Global state interactions* (pp. 293–312). Cambridge, MA: MIT.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing* (pp. 67–82). Hillsdale: Erlbaum.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gross, H.-M., Heinze, A., Seiler, T., & Stephan, V. (1999). Generative character of perception: A neural architecture for sensorimotor anticipation. *Neural Networks, 12*, 1101–1129.

Haag, J., & Borst, A. (1998). Active membrane properties and signal encoding in graded potential neurons. *The Journal of Neuroscience, 18*(19), 7972–7986.

Hall, Z. W. (1992). *Molecular neurobiology*. Sunderland: Sinauer.

Haugeland, J. (1998). *Having thought*. Cambridge, MA: Harvard U. Press.

Hertz, L., & Zielker, H. R. (2004). Astrocytic control of glutamatergic activity: Astrocytes as stars of the show. *Trends in Neurosciences, 27*(12), 735–743.

Izhikevich, E. M. (2001). Resonate and fire neurons. *Neural Networks, 14*, 883–894.

Izhikevich, E. M. (2002). Resonance and selective communication via bursts in neurons. *Biosystems, 67*, 95–102.

Izhikevich, E. M. (2007). *Dynamical systems in neuroscience*. Cambridge, MA: MIT.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*(11), e1000209. doi:10.1371/journal.pcbi.1000209.

Koziol, L. F., & Budding, D. E. (2009). *Subcortical structures and cognition*. New York: Springer.

MacKay, D. M. (1956). The epistemological problem for automata. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 235–251). Princeton: Princeton University Press.

MacKay, D. M. (1969). *Information, mechanism and meaning*. Cambridge, MA: MIT Press.

Marder, E. (2012). Neuromodulation of neuronal circuits: Back to the future. *Neuron, 76*, 1–11.

Marder, E., & Thirumalai, V. (2002). Cellular, synaptic and network effects of neuromodulation. *Neural Networks, 15*, 479–493.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition*. Dordrecht: Reidel.

McDermott, D. (1981). Artificial intelligence meets natural stupidity. In J. Haugeland (Ed.), *Mind design* (pp. 143–160). Cambridge, MA: MIT Press.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254–278.

Nauta, W. J. H., & Feirtag, M. (1986). *Fundamental neuroanatomy*. San Francisco: Freeman.

Nedergaard, M., Ransom, B., & Goldman, S. A. (2003). New roles for astrocytes: Redefining the functional architecture of the brain. *Trends in Neurosciences, 26*(10), 523–530.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton.

Newman, E. A. (2003). New roles for astrocytes: Regulation of synaptic transmission. *Trends in Neurosciences, 26*(10), 536–542.

Nieuwenhuys, R. (2001). Neocortical macrocircuits. In G. Roth & M. F. Wullimann (Eds.), *Brain evolution and cognition* (pp. 185–204). New York: Wiley.

Perea, G., & Araque, A. (2007). Astrocytes potentiate transmitter release at single hippocampal synapses. *Science, 317*, 1083–1086.

Pezzulo, G. (2008). Coordinating with the future: The anticipatory nature of representation. *Minds and Machines, 18*, 179–225.

Pezzulo, G., Candidi, M., Dindo, H., & Barca, L. (2013). Action simulation in the human brain: Twelve questions. *New Ideas in Psychology.* http://dx.doi.org/10.1016/j.newideapsych.2013.01.004

Powers, W. T. (1973). *Behavior: The control of perception*. Chicago: Aldine.

Ransom, B., Behar, T., & Nedergaard, M. (2003). New roles for astrocytes (stars at last). *Trends in Neurosciences, 26*(10), 520–522.

Roberts, A., & Bush, B. M. H. (Eds.). (1981). *Neurones without impulses*. Cambridge: Cambridge University Press.

Roesch, E. B., Nasuto, S. J., & Bishop, J. M. (2012). Emotion and anticipation in an enactive framework for cognition (response to Andy Clark). *Frontiers in Psychology, 3*, 1–2.

Slezak, M., & Pfreiger, F. W. (2003). New roles for astrocytes: Regulation of CNS synaptogenesis. *Trends in Neurosciences, 26*(10), 531–535.

Sokolov, E. M. (1960). Neuronal models and the orienting reflex. In M. Brazier (Ed.), *The central nervous system and behavior* (pp. 187–276). New York: Josiah Macy Jr. Foundation.

Tani, J., & Nolfi, S. (1999). Learning to perceive the world as articulated: An approach for hierarchical learning in sensory-motor systems. *Neural Networks, 12*, 1131–1141.

Varela, F. J. (1979). *Principles of biological autonomy*. New York: North Holland.

Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition, 34*, 72–87.

Verkhratsky, A., & Butt, A. (2007). *Glial neurobiology*. Chichester: Wiley.

Viggiano, D., Ibrahim, M., & Celio, M. R. (2000). Relationship between glia and the perineuronal nets of extracellular matrix in the rat cerebral cortex: Importance for volume transmission in the brain. In L. F. Agnati, K. Fuxe, C. Nicholson, & E. Syková (Eds.), *Volume transmission revisited* (Progress in brain research, Vol. 125, pp. 193–198). Amsterdam: Elsevier.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences, 1*, 97–125.

Chapter 17
General Homeostasis, Passive Life, and the Challenge to Autonomy

Stefano Franchi

Abstract The paper argues that the conception of life as generalized homeostasis developed by W.R. Ashby in *Design for a Brain* and his other writings is orthogonal to the traditional distinction between autonomy and heteronomy that underlies much recent work in cellular biology, evolutionary robotics, ALife, and general AI. The distinction is well-entrenched in the Western philosophical canon but it fails to do justice to Ashby's conception of life. We can assess the philosophical and technical viability of the general homeostasis thesis Ashby advocated, the paper argues, through the construction of virtual cognitive agents (i.e. simulated robots in a physically plausible environment) that replicate the architecture of Ashby's original homeostat through a CTRNN-like network architecture, whose outline implementation is then discussed.

Keywords Homeostasis • W.R. Ashby • Life • Autonomy • Heteronomy

17.1 Autonomy, Body, and Mind

The fundamental understanding of autonomy and heteronomy and their respective alignments with other important concepts (body vs. mind, animal vs. human, passions vs. reason, etc.) has hardly changed throughout the centuries. These distinctions are crucial to our understanding of living beings as well as to our efforts at designing artifacts that replicate simplified versions of their functions. Recent work in cellular biology, evolutionary robotics, A-Life, and general AI (Varela 1979; Di Paolo et al. 2010; Pfeifer and Bongard 2007) allows us to reexamine these classic oppositions from a fresh new perspective. Even more importantly, as I will argue below, technical work coming out of the cybernetics and AI traditions provides the means for a re-evaluation of the classic categories that may produce a wholly novel perspective on life itself. Before we can examine what such a perspective may be, though, it is necessary to introduce a precise definition of the *philosophical* concepts of autonomy and heteronomy.

S. Franchi (✉)
Texas A&M University, College Station, TX, USA
e-mail: stefano@tamu.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_17

In the technical literature, "autonomy" has a very broad meaning. In a widely adopted textbook, for instance, George Bekey (2005), defines it as the capacity to operate in the real word without any form of external control. An autonomous robot, therefore, is a device that can operate without continuous and presumably human supervision. As Bekey indicates, living beings are the canonical examples of autonomous systems, since they are capable of maintaining their internal structures, scout the environment to locate material for sustenance, and, within limits, adapt to environmental change. "Autonomy" is almost a synonym of "living." In the disciplines of robotics and AI, the limited degree of autonomy of our current robots is simply an indication of our still inadequate understanding and implementation of autonomy. A fully autonomous artificial system would thus be formally indistinguishable from a natural living system. As Bekey states, "living systems are the prototype of autonomous systems": they are the original (proto-) form of autonomy that scientific and engineering discipline strive to functionally replicate. As a consequence, the concept of heteronomy has no place in the technical literature. We are certainly well acquainted with artifacts that require continuous supervision—cars and drones as well as traditional industrial robots belong squarely in this category—but they are usually called "non-autonomous" rather than "heteronomous." In other words, "autonomy" is a property that artificial systems may or may not have whereas all living systems do.

The philosophical conception has a narrower scope and a more precise meaning. Far from being synonymous with "life," autonomy identifies a strict subset of living beings. An autonomous system is a system capable of giving itself laws that govern its own behavior. Heteronomy, on the contrary, designates an entity's inability to do the same, its behavior being ruled by an external agency or law (*nomos*). In other words, only a few living systems are rightfully called "autonomous": namely those, in Kantian terms, that posses a capacity for self-legislation. All other forms of life are heteronomous. Although the terminology is Kant's, the distinction is much older. Plato and Aristotle already distinguished between forms of life that could and could not govern themselves. Most of the subsequent philosophical tradition, with a few scattered exceptions (e.g. the Cynic school, La Mettrie's *L'Homme-machine*, etc.) has followed them quite closely.[1] Kant sums up the classical view by explicitly opposing the passive receptivity of the body to the active and spontaneous role of rational subjectivity:

> It is not from *receptivity*, but from the *spontaneity of the subject* that the aggregate of perception becomes a system, thus from that which the understanding makes out of this simple material, hence *autonomously* not *heteronomously*. (1993, 22:447/165)

[1]Classical *loci* of this distinction are the so-called "function of man argument" in book I of Aristotle's *Nichomachean Ethics* Aristotle (1984, v. 1, 1097b22–1098a18) and the argument for the tri-partite composition of the psyche in *Phaedrus* (Plato 1997, 237c–241d) and *Republic* (Plato 1997, 435e–441c). Both Aristotle and Plato stress that animal life is propelled by "drives" (or "passions"). The structure of this argument is repeated almost verbatim in Descartes's study of the passions (see Descartes (1988), esp. articles 7–37 of Part 1).

This distinction was usually mapped onto the other oppositions I mentioned above: body/mind, passions/reason, human/animal. The argument is simple and was often presented in these terms: when an outside agency (the world) impinges upon an entity's *body*, its necessary reactions (its *passions*) force it to act accordingly. The passions are literally the "affections" of the body (*ta pathémata*) that allow it to be in contact with the external world by producing some internal change as a response to external stimuli. Since the body, in turn, will initiate some reaction as a response to its affections, it follows that its behavior is always passion-mediated and hetero-directed. The rule of its behavior will always come from the "other," for its whole range of actions will ultimately amount to a set of responses to external stimuli. The absence of external control that was a crucial component of the technical definition of autonomy is quite irrelevant in the philosophical meaning of the term. Consider a functionally simple living being such as the bacterium *E. coli*. This organism is capable of producing a sophisticated set of behaviors: it can locate food sources by actively swimming up favorable chemical gradients and away from poisonous ones, it can repair its internal structure by replacing parts of its own body with metabolized components drawn from the environment, it can successfully reproduce, and so forth. Yet, all these behaviors—which, as a whole, are still well beyond the capabilities of current artificial autonomous systems—are always initiated from the outside: they are always reactions to external stimuli. According to the classical conception, they are perfectly heteronomous. Even the fairly complex mechanisms that govern *E. coli*'s capacity to move around and successfully forage for food (Berg 2004) are ultimately initiated by an externally generated "passion": the stimulus that results from the organism's receptivity to its environment.

Only an organism endowed with an internal agency capable of resisting the body's passions (mind's reason) could resist the outside world and develop its own self-directed behavior. Mind-endowed and therefore rational humans are the only entities capable of doing so, and thus the only true autonomous beings. Here again, Kant's formulation encapsulate this series of oppositions by explicitly stating that human (or, rather, rational) life (all of it, not just moral life) is not immediately connected to the outside world (via sensuous receptivity). Instead, it is detached from it: humans' connection to the world is mediated and hence governed by rules that they (spontaneously) give to themselves. The first maxim of common human understanding, Kant states in the third *Critique*, is

> to think for yourself [which] is the maxim of freeing oneself from prejudice [...] the maxim of a *never passive reason.* The inclination toward the latter, and therefore toward the *heteronomy of reason*, is called prejudice. (1952, AK 5.294)

Thus, the classical distinction produces a double series of oppositions:

- *Autonomy* → Reason → Mind → Activity → *Freedom* ⇒ Humans
- *Heteronomy* → Passions → Body → Passivity → *Slavery* ⇒ Animals

Figure 17.1 summarizes the structural configuration that these chains of oppositions assume in the predominant strand of the Western philosophical canon, on the basis of the early definition in *Nichomachean Ethics*. As Aristotle states: "Human

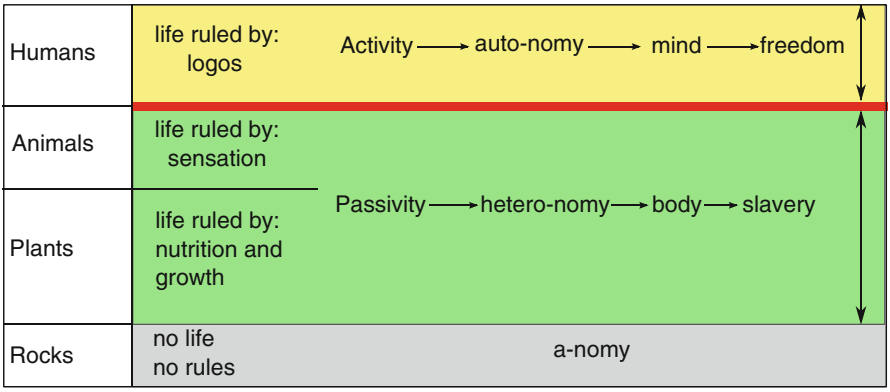


Fig. 17.1 Aristotle's classification of different forms of life according to the principle ruling their *characteristic* behaviors

life's mode of operating (*érgon anthrópou*) is an activity of their internal living principle (*enérgeia psuchés*) pursued in accordance with the rational principle (*katà lógon*) or at least not independent of it."[2] Notice that forms of life listed higher up in the table include those below as their necessary and yet non-characteristic components. All living beings must feed themselves, grow, and reproduce and these fundamental need define the the basic drives of hunger, self-preservation, and sex. Some living beings are just that and their lives are completely ruled by drives. In Aristotle's terminology, they are "plants." Some living beings go beyond basic drives, yet they include them as part of their more elementary behaviors. An animal is also a plant, to a limited extent: it will behave like one when its life is determined by the basic drives. Yet its characteristic, animal-only, behaviors will be ruled by sensation (or by the passions). Similarly for humans, who are also plant-like and animal-like, yet *not characteristically so*. A direct consequence of the classical view is that human life—and, to a lesser extent, animal life—is traversed by an intrinsic contradiction, as Jonas will say. Humans are *characteristically* rational autonomous beings, but they are *also* passionate heteronomous animals and nutrition- and reproduction-oriented plants. The main challenge of human life, as Aristotle and Plato had already stated and as Kant will repeat, is how to reconcile these different "souls" that inhabits their being in order to be truly human. Plato's *Phaedrus* epitomized this view with the image of human existence as a carriage being drawn by two unruly horses (passions and drives) led by a charioteer who is always struggling to control them (reason). On the contrary animals, and especially plants, have it easier: their fully heteronomous existence removes any conflict between spontaneously and internally generated goals and externally generated pressures. Animals and plants are both machine-like beings whose whole life is

[2] *Eth. Nic.* I,7, 1098a7: "ἔργον ἀνθρώπου ψυχῆς ἐνέργεια κατὰ λόγον ἢ μὴ ἄνευ λόγου."

an untroubled journey completely governed from the outside. I would like to stress once again that Kant's notion of autonomy (which is accepted by the enactivism movement, to which I will come back later) is much stronger than the concept commonly used in standard Cognitive Science and AI. Consider a human diving into a pool. From a Cognitive Science perspective, it would be a quintessential example of autonomous behavior: the human came up with a goal—"dive into a pool"—and then executed a series of actions—a "plan"—whose final result was the achievement of the initial goal. From a Kantian perspective, however, we cannot say if the observed behavior is autonomous or heteronomous unless we know *why* the human dove into the pool. Was the human responding to a feeling of excessive heat? Then the behavior was *heteronomous* by definition, since it was the effect of a bodily passion ultimately produced by the environment (or, more precisely, by the almost automatic interaction between the body's receptors and actuators and the environments). On the contrary, if the human's dive was prompted by the decision to save someone from drowning, then it was (probably) *autonomous*.[3]

The structural configuration reproduced in Fig. 17.1 produces an obvious dualism between the autonomous-human and the heteronomous-animal regions of the chart. We are perhaps best acquainted with the form this dualism took in Descartes, who assigned two different substances (*res cogitans* and *res extensa*) to the autonomous and heteronomous regions and described their mode of functioning as, respectively, non-mechanical or free and fully mechanical or watch-like. Yet, the Cartesian dualistic ontology is just one instance of the numerous debates and negotiations throughout the history of philosophy and, more recently, of biology and cognitive science. The main focus of these confrontations has always been the exact positioning of the horizontal line that partitions the domain of living beings into two non-overlapping subsets. Setting themselves in self-conscious opposition to the traditional philosophical viewpoint, some schools of thought (from the eighteenth century chemical mechanicism championed by La Mettrie and D'Holbach, to nineteenth century's reflexology and psychoanalysis, up to the recent school of evolutionary psychology) have tried push the boundary between autonomy and heteronomy *upward*. In other words, they have tried, by different means, to extend the scope of the heteronomous region to include more and more organized forms of life such as human life. As a consequence, the scope of the autonomous region traditionally identified with autonomy was shrunk to almost nothing. By increasing the role of the body and of its affections or passions in human and non-human behavior and turning all cognitive functions into bodily (and, ultimately, environment-directed) mechanisms of different kinds, these traditions sought to eliminate the classic dualism at the root. The necessary trade-off was a dramatic

[3]I am skipping over a few important details here, in order to get the main point across. A fuller analysis on Kantian grounds would have to specify the exact reasoning that led the human to dive. Kant himself provided the blueprint for all such analysis of autonomy in the well-known example of the shopkeeper he discusses in the *Groundworks* (1999, AK 397ff.). Hans Jonas (1966, 115ff.) built directly on this argument in his rebuttal of early cybernetics' claim that a self-driven target-seeking torpedo would be acting autonomously.

reduction of the scope of autonomous agency, a necessary consequence that was perhaps most apparent in Freudian psychoanalysis, but is similarly foregrounded in evolutionary psychology.

Instead, the work in ALife, evolutionary robotics, and biology I mentioned above follows the reverse strategy. Instead of pulling the divide between heteronomy and autonomy up, it pushes it *down*. It seeks to expand the scope of autonomy by moving it away from its traditional confines of higher level, mind-only (e.g. "rational") cognitive functions to include lower-level biological functions. Whereas reflexologists, psychoanalysts, and evolutionary psychologists (to name just a few examples) were happy to stress the heteronomous functioning of the mind, recent emphasis on autonomous behavior in cognitive science, the philosophy of mind, and evolutionary robotics has identified the principle of philosophical autonomy in the most elementary forms of life. The best known example of this trend is perhaps Jonas's (1966) view of any form of metabolism, from bacteria "upward," as containing the essential kernel of philosophical freedom in Kant's sense: the possibility to establish the rules governing one's own behavior. In a later essay that summarizes the argument he presented at length in *The Phenomenon of Life*, Jonas explicitly lists almost all the terms we saw at work in the canonical philosophical series—autonomy and heteronomy (or dependence), freedom and slavery (or necessity), and so on—and affirms their presence in all forms of life:

> The great contradiction that man discovers in itself—freedom and necessity, autonomy and dependence, ego and world, connectedness and isolation, creativity and mortality—are present *in nuce* in life most primitive forms, each of which maintains a precarious balance between being and nonbeing and from the very beginning harbors within itself an inner horizon of "transcendence." [...] we maintain that metabolism, the basic substratum of all organic existence, already display freedom—indeed that it is the first form freedom takes. (1996, p. 60)

Jonas's view, in turn, has played an important role in the theory of life as autopoiesis (Varela 1979; Weber and Varela 2002; Barandiaran and Ruiz-Mirazo 2008) which constantly points to the intrinsic autonomy of the cell and turns the body into the locus of autonomy.

The symmetry between the two anti-dualism strategies I sketched shows that they actually share the traditional philosophical articulation of both autonomy and heteronomy. What is really at stake in both strategies is the mapping of that distinction over the range of bodily and cognitive functions in humans and non-humans alike. The general configuration of the problem, however, remains untouched and the chain of oppositions headed by, respectively, autonomy and heteronomy is not challenged by either school. Mind, reason, freedom, activity, and spontaneity always go together as well as do body, passions, passivity, and slavery (or dependence). The former are the markers of autonomy, the latter signal heteronomy. We could even assert that Freud's fully heteronomy-centered position, for instance, is the flip side of Jonas' fully autonomy-centered theory. While disagreeing on how to partition actual living beings between the two received categories, both presuppose the same canonical characterization of life-forms .

The recent history of cybernetics offers a significant exception to this symmetry: the work of W. R. Ashby, which sits awkwardly in this debate as it seems to participate in both movements. On the one hand, Ashby quotes approvingly Pavlov's and Freud's work. In an early article (1954, p. 122), while not mentioning Freud directly, Ashby suggests an explicit comparison between "the basic instinctual forces seen by the psychoanalyst" and the "essential variables seen by the mechanist," namely the variables he himself had seen and modeled in *Design for a Brain*, whose first edition was published just 2 years earlier. In a private note, he even stated that "at the important times [of my life], in the words of Freud, I do not live but 'am lived'."[4] Ashby's comparison of the basic components of his cybernetic model of the brain to psychoanalysis' "instinctual forces" (i.e. *Triebe* such as sex and hunger) suggests that he shared Freud's attempt to show that human life is, in classical terms, essentially heteronomous. Life's most important events are determined by unconscious drives that reason's autonomously generated plans have no control over. Ashby's notebook's remark confirms this interpretation.

On the other hand, the followers of the theory of autopoiesis have successfully appropriated Ashby's work as supporting Jonas's view of all forms of life as essentially autonomous (Di Paolo 2005; Egbert et al. 2010; Ikegami and Suzuki 2008). Di Paolo et al. (2010) explicitly state that the enaction paradigm and its five, mutually implicating, foundational concepts—*autonomy*, sense-making, emergence, embodiment, and experience—require an Ashbian model of regulation in order to be effectively implemented in natural and artificial living systems. Autonomy, at least in its Jonas-derived enactive reformulation, depends upon Ashby's theory of homeostatic adaptation to the environment.

We may wonder how a follower of the heteronomy-centered anti-dualist strategy could become a champion of autonomy. My claim is that Ashby's model of cognition is actually orthogonal to both autonomy and heteronomy and points toward a genuine, yet very much underdeveloped third way that provides an alternative conception of life. To put it differently, Ashby's model actually does what neither Freud nor Jonas attempted: it breaks the conceptual chains that link autonomy to mind and reason and heteronomy to body and passions. This is why his work may be aligned with either camp—because it really sits in neither.

17.2 General Homeostasis

In *Design for a Brain*, Ashby holds that homeostatic adaptation to the environment governs *all* aspects of *all* forms of life. This proposition—which I will call, following Ashby, the *general homeostasis thesis*—holds that we can interpret complex

[4]This passage from Ashby's notebooks is reported by Andrew Pickering (2010, p. 112) and discussed by Helge Malmgren (2013) in his response to Franchi (2013). I discuss some of the possible relationships between cybernetics and psychoanalysis and give a very brief review of the sparse literature on the subject in Franchi (2011a).

behavior as the visible manifestation of a complex system running to equilibrium (1960, 55–56). The thesis succinctly represents Ashby's fundamental claim about the behavior of *any* complex system—be it a physiological, psychological, or even a social one. On the one hand, Ashby seems to be pushing heteronomy all the way up: all kinds of purposeful behavior, from phototaxis to chess-playing (1952a), are the results of processes running to equilibrium—i.e. adapting to environmental conditions—and are therefore hetero-directed. On the other hand, we may argue that generalized homeostasis actually undercuts the heteronomy/autonomy opposition. The homeostat, the prime model of a general homeostatic organism is really a "sleeping machine" of sorts (Walter 1961, 111) that will go to extraordinary lengths to go back to sleep—i.e. to equilibrium; its ever possible failure to reach equilibrium will bring about its demise. The Ashbian organism will try to accommodate itself to its environment by whatever means necessary: it is essentially a *passive* machine whose activity is a by-product of its search for non-action. It is also a *contingent* machine insofar as its "search" for equilibrium will involve essentially random processes, as Ashby never tired to stress. Under general homeostasis, all life forms (life itself) turn out to be made out of contingent, dispersed, embedded, equilibrium-seeking beings that may (or may not), among the many devices they use to achieve quiescence, include the generation of goals and their related consciousness.

Ashby's stress on inactivity as life's primary goal throws a wrench into the autonomy/heteronomy distinction, as Jonas himself had clearly seen (1984, 62). A (traditionally conceived) autonomous being is the source of its own actions or, to be more precise, of the rules presiding over those actions. On the contrary, a heteronomous being is at the receiving end of someone else's (or something else's) actions. The common link between the two views lies in the primacy of action: the traditional view of life, from Plato to Jonas, includes only autonomous leaders or heteronomous followers, so to speak. But homeostatic life is neither, and this is the feature that makes it an appealing candidate for a third way that cuts across the traditional autonomy/heteronomy opposition entrenched in the Western tradition, which critics of the dualism it produces never really challenged. Ashbian life follows (it is heteronomous), but it makes up its own rules when doing so (it is autonomous). And yet it makes its rules (like an autonomous life form), only when the environment forces it to (like heteronomous forms of life). No wonder the first witnesses to the homeostat's performances could hardly wrap their head around it (von Foerster et al. 1953, 95).

While the participants in the debate upon the applicability of autonomy and heteronomy to cognitive and biological functions could rely upon concepts that were at the very least a few centuries, if not a few millennia old, the homeostatic outlook on life did not enjoy such a strategic advantage. We have known the outline of the answer to a question such as "what it is like to live an autonomous life?" at least since Plato and we have come to know it better and better with all subsequent philosophical elaborations. We only need to determine the important details, perhaps down to the cellular level, while we struggle to understand *who*, exactly, can be said to live it. Otherwise put: we know what a human-like "autonomous life" is. We only need to decide if anyone is living it, or if it is a self-reassuring figment of our

imagination (as Freud held). And similarly for heteronomy: we know what a plant-like, drive-determined, and an animal-like passion-driven life is. We only need to determine if any living being is living it or if it is just an empty concept that arose as a by-product of our mistaken infatuation with the exact sciences, as Jonas claimed (1966, pp. 64–98).

When we face the similar question "What it is like to live a passive/contingent, homeostatic life?" the answer is not so easy. Neither Ashby nor his contemporary or recent followers could provide a systematic exploration of homeostatic life. On the one hand, the evident technological limitations of Ashby's simulating device, the homeostat, hampered a full investigation. On the other hand, Ashby's followers have preferred to focus on the *technical* aspects of his theory (e.g. the concept and mechanisms of ultrastability) while leaving in the background the more philosophical aspects of the general homeostasis thesis. Indeed, a widely shared view about Ashby's work holds that he borrowed his conception of homeostasis from the physiologists who worked out the concept of organism/environment equilibrium between the end of the nineteenth century and the beginning of the 20th. His main contribution would be limited to the formal description of the mechanisms of homeostasis he provided in *Design for a Brain*. In my opinion, this view ignores an essential component of his work. While it is true that—from the theoretical point of view—Ashby merely extended the *scope* of homeostasis within the the organism, the generalization of homeostasis produced a novel *philosophical* interpretation of life that was not only fully absent from his contemporary physiologists' work, but even in contradiction with it. A brief discussion of the relationship between Ashby's and physiology's homoestasis will confirm this point.[5]

Ashby lifted the term mostly from Walter Cannon, whose successful 1932 book, *The Wisdom of the Body*, had systematized the theory—which Claude Bernard advanced first and then later Ernest Henry Starling and John Scott Haldane extended—by supplementing it with additional experimental evidence.[6] But Ashby's theory of the role played by homeostasis in the overall econoy of the organism—and especially in humans—is very different from his predecessors'. Bernard held that "the constancy of the internal environment [i.e. homeostasis] is the *condition* for free and independent life" (1974, p. 84, emphasis added). In other words, homeostatic regulations take care of the heteronomous layer of life in order

[5]For a more detailed discussion of Ashby's relationships with his predecessors, the cybernetics movement in general, and the relationship between physiological and cybernetic homeostases, see Franchi (2011b).

[6]Walter Cannon coined the term in 1929, although Claude Bernard had introduced the fundamental idea about 50 years earlier (1966[1878–1879]). Next, Lawrence Joseph Henderson in the United States and John Scott Haldane in Britain (Henderson 1928; Haldane 1917 and especially Haldane 1922) popularized it in the English speaking world. The title of Cannon's celebrated 1932 book, *The Wisdom of the Body*, pays homage to William Starling's homonym lecture, which had dealt with the self-compensating mechanisms of the heart (Starling 1923). As Pickering (2010, fn. 12, pp. 422–424) makes clear, Ashby elaborated his theory well before he became familiar with Cannon's work, although he only started to use the term "homeostasis" after reading it.

to let the autonomous ("free and independent") component rule. William Cannon is even more explicit. In the final chapter of *The Wisdom of the Body* he states explicitly that bodily homeostasis gives the organism the freedom for "the activity of the higher levels of the nervous system [. . . and. . .] for its more complicated and socially important tasks" (Cannon 1939[1932], p. 302). Cannon sees bodily homeostasis as the precondition for higher and substantially *non-homeostatic* processes. The view had already been anticipated by John S. Haldane within an explicitly Kantian framework when the British physiologist stated that "it is unmeaning to treat consciousness as a mere accomplishment to life or to ignore the differences between *blind organic activities and rational behavior*" (1917, p. 115, my emph.). Like Aristotle, Descartes, or Kant, the early twentieth century physiologists held on to the view that homeostasis explains only the lower level of behavior: it is a precondition that will be overruled by the higher, non-homeostatic ones. The vegetative and emotional parts of the soul, as Aristotle had stated, are subordinated to the rational part.

On the contrary, Ashby self-consciously promoted a far more radical view: homeostatic regulation, he claimed, rules all aspects of behavior, from the lower-level physiological exchanges with the environment to higher level cognitive functions. In *Introduction to Cybernetics*, Ashby makes this point clearly. While Walter Cannon had treated adequately the subject of bodily regulation, he states, his own goal is to write the "much larger book" that would show how "all the organism's exteriorly-directed activities—its "higher" activities—are all similarly regulatory, i.e. homeostatic" (1956, pp. 195–196). The next few lines in the text make also clear that he believed his previous work—*Design for a Brain*—was such a book, or at the least the beginning of it. In short: by extending the scope of homeostatic regulation to the whole complex of an organism's functions—hence the moniker "generalized homeostasis" he applied to his own theory—Ashby produced a genuinely different philosophical understanding of life that put him at odds with the received philosophical tradition still underwriting the physiologists' work.

I propose to get a better understanding of Ashby's broader claim by focusing on his original device as a practical means toward a further exploration of the view of the passive contingent life he advocated. Putting back at the center of our theoretical attention the conceptual simulation of Ashbian life as general homeostasis means to start from a modern version of Ashby's homeostat. In a research program consistent with the previous discussion, the first step would be the replication of Ashby's original setup. Although implemented in software rather than in the electro-mechanical hardware Ashby worked with, the replicated homeostat would strive to be functionally equivalent to the original. The second step is the reproduction of some of the experiments that Ashby himself detailed in *Design for a Brain*. Third, we would conduct some preliminary investigations of the general characteristic of the homeostat. Finally, and most importantly, by embedding homeostats in sensor- and motor-endowed "virtual creatures" acting in a real "environment"—virtual robots or, to put it differently, virtual minimally cognitive agents (Beer 1996)—the research would expand the scope of Ashby's simulations to other, open-ended domains (i.e. with "real" environments). In spite of the renewed interest in Ashby's

work, especially in the wake of Di Paolo's work since 2000, there have been no attempts to provide digital simulators of Ashby's homeostat that preserve all the features of his model, including the network's dynamics and the completely stochastic selection of alternative strategies that he saw as theoretically crucial. In the 1960s, there were a couple of failed efforts to extend Ashby's work (Haroules and Haire 1960; Wilkins 1968), and a limited batch-oriented digital simulator (Capehart and Terry 1968). Even the most recent models (from Di Paolo 2000 to Izquierdo et al. 2013) decouple ultrastability from Ashby's larger thesis about generalized homeostasis and end up with systems which are theoretically at odds with his original intuition.[7]

I have been working on the research program I just sketched since 2011, and I am currently focused on the last phase. Namely, the extension of Ashby's general homeostatic model to virtual open-ended robotic domains. It is perhaps worth stressing that while Ashby claimed repeatedly that the structure of his device's abstract *performances* bore some similarities with living beings' actual behavior, he never claimed (nor could he reasonably do so) that the *device* itself had any biological plausibility. The homeostat's lack of biological reality, however, is actually an advantage for the research I am pursuing. It allows the researcher to focus on its abstract features while leaving in the background important, yet premature questions about the possible biological realizations of general homeostasis. For these methodological reasons, and not for mere archaeological interest, the suggested replica of the homeostat will follow as closely as possible Ashby's original implementation. In the remainder of the chapter I will describe briefly the suggested implementation and the experiences carried out on it so far.

Let me begin with the basic structure of the homeostat and its simulated replica. Ashby's device was made up of four inter-connected units (Fig. 17.2) each one consisting of a needle dipped in an electrically conductive trough and connected to a magnet. The position of each of the needles can deviate either side of a central point. The positions of the magnets represent the essential variables that the system must keep within bounds. All units are connected to each other. The *torque* that operates on each magnet is approximately proportional to the sum of its own output current plus the currents that the other three units produce. The current that a unit sends out and then feeds back to itself is proportional to the deviation of the magnet from its central position, as the needle picks up a potential dependent on its position in the trough. The viscosity of the conducting medium dampens the system and makes it more stable. Following Ashby's own mathematical analysis (1960, pp. 246ff.), the 4-unit homeostat is modeled by a N-units network whose individual units j are dynamic devices with values described by the equation

$$m\ddot{y}_j = -v\dot{y}_j + \sum_{i=1}^{N} w_{ij} y_i$$

[7] See Franchi (2013) and more extensively Franchi (2011b) for a discussion of this point.

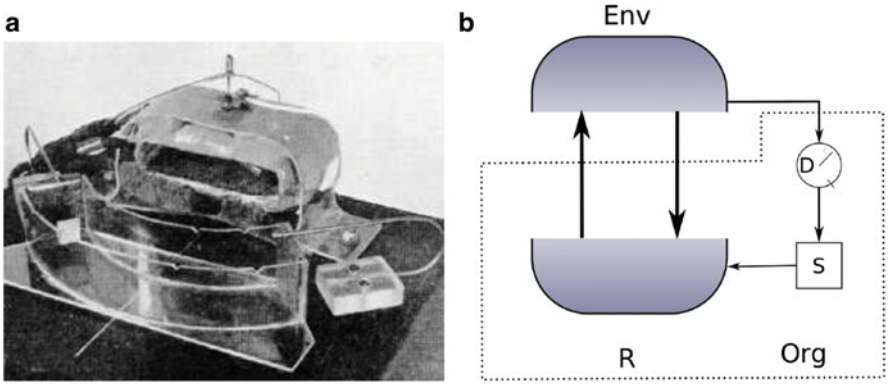


Fig. 17.2 The original homeostat from *Design for a Brain* (1960, p. 101). Ashby's well-known original device, a fully-connected collection of four double-feedback units each including a random step mechanism S capable of altering the weight of all its connections when the unit's value exceeds a preset value measured by the sub-device D. (**a**) The needle-magnet part of a homeostatic unit. (**b**) A homeostatic unit's functional diagram

where v is the viscosity of the medium the unit's needle works in (the original trough), m depends on the unit's needle's mass, and w_{ij} is the weight of the connection from unit i to unit j. Mechanical analogies aside, m obviously represents a unit's inertia and it is similar to a neuron's bias, while v is a damping factor. A single, negatively self-connected unit (Ashby's base case) turns into a harmonic oscillator, since the summation term reduces to $-w_{jj}y$. The oscillator is damped if viscosity $v > 0$. Thus, in the general case, a homeostat is a network of interconnected damped harmonic oscillators. This model is structurally similar to a CTRNN network (Beer 1995), with the important difference that the differential equation describing the nodes' values is analytically solvable and does not require the use of a numerical solver for its software implementation. In addition to its basic behavior as described by the equation above, each unit periodically checks at interval T that its value has not exceeded a critical value V, in which case it randomly resets the weights of all of its connections (except its self-connection, which is always negative) with a new value in the interval $(-1, 1)$. Thus, the behavior of each node in a network of size N is completely described by the set of 4 parameters: $\{m, v, V, T\}$ plus N weights, for a total of $N + 4$ values. These are the values to be used in the genetic algorithm simulations based on the model.

I have tested this model on the original experiments Ashby described in his major work. As an example, Fig. 17.3 reproduces the trace from a run of *Design of a Brain*'s second experiment, which shows the homeostat's ultrastability in action. Three units (roughly interpretable as Agent-Motor, Agent-Sensor, and Environment) are connected circularly ($M \rightarrow E \rightarrow S \rightarrow M$). M and S are self-regulated, E is not. The system is originally stable. At A, the value of M is manually decreased, provoking a local instability and resulting in oscillating behavior until equilibrium is restored. At B, the polarity of the $E \rightarrow S$ connection is switched (as if a

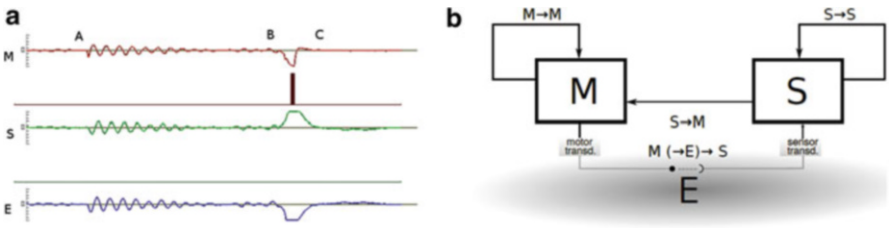


Fig. 17.3 A replica of Ashby's second experiment in *Design for a Brain* (p. 106) with the software replica of the Homeostat described in the text. (**a**) Trace from 3-unit simulation. (**b**) Functional diagram showing Ashby's "dispersion"

sensory transducer's mode of functioning had been reversed) and a downward manual displacement of *M* is repeated. The system becomes unstable until three actions of the stepping mechanism restore stability at *C*. From the biological point of view, we can interpret the setup just described as an extremely simplified model of an elementary organism trying to achieve equilibrium with respect to an external stimulus, like, for instance the form of chemotaxis we observe in bacteria, or a very simple form of phototaxis. At the more abstract level, it can be seen as a more sophisticated, homeostasis-based and ultrastable implementation of a type-1 "vehicle" as those described by Valentino Braitenberg in his classic monograph (1984).

As outlined in the final step of the research program I sketched above, the homeostatic model I just described has then been used as the controller of a simple Khepera-like robot in a physically plausible virtual robotic environment (WEBOTS, Michel 2004). Depending on the polarity of the stimulus affecting the robot, which is hardware-dependent and, biologically speaking, reflects the physiological features of the sense-organ it models, the robot is expected to achieve equilibrium with respect to the stimulus source (a randomly positioned light source) by either running away from the stimulus or moving toward it. The simple experiments carried out so far produce the expected behaviors.

These preliminary results are encouraging insofar as they confirm Ashby's original intuitions: they provide initial, although still tentative evidence that we may conceive life as a passive-contingent phenomenon, to use the terminology I introduced earlier, which is neither autonomous nor heteronomous, but, paradoxically enough, both at the same time. In order to move from mere plausibility toward a more substantive view, the next research phase involves extending the experiments to creatures engaged in different forms of mono- and multi-sensory taxis behaviors such as binocular light-seeking while testing the systems' robustness under different kinds of bodily lesions (Di Paolo 2000). Of particular interest is also the simulation of an elementary multi-tasking goal involving two non-necessarily cooperative sub-tasks such as, for instance, a simplified form of obstacle-avoidance carried out together with a taxis behavior, or, alternatively, forms of taxis involving multiple sensory modalities. Ashby was working on such an extended simulator

in the years after the publication of his major work (DAMS, *Dispersive and Multistable System*, (Ashby 1952b, p. 171)). His efforts ultimately failed (Husbands and Holland 2008, pp. 125–126), thereby establishing the homeostat's reputation as a technically resourceful yet ultimately non-viable cognitive architecture. We may well wonder whether the homeostat's failure, and, perhaps, the failure of Ashby's overall philosophical project to ground life upon homeostasis was due to the technical limitations of 1950s simulation technology or whether the homeostat represents a concrete, although minimal, embodiment of an alternative conception of life that is neither autonomous nor heteronomous. The answer to this question is predicated upon our building and testing fully general homeostatic entities operating in settings of ever increasing complexity.

References

Aristotle. (1984). *The complete works of Aristotle*. Princeton: Princeton University Press.

Ashby, W. R. (1952a). Can a mechanical chess-player outplay its designer? *The British Journal for the Philosophy of Science, III*(9), 44–57.

Ashby, W. R. (1952b). *Design for a brain* (1st ed.). New York: Wiley.

Ashby, W. R. (1954). The application of cybernetics to psychiatry. *The British Journal of Psychiatry, 100*(418), 114–214.

Ashby, W. R. (1956). *An introduction to cybernetics*. London: Chapman and Hall.

Ashby, W. R. (1960). *Design for a brain* (2nd ed.). New York: Wiley.

Barandiaran, X., & Ruiz-Mirazo, K. (2008). Modelling autonomy: Simulating the essence of life and cognition. *Biosystems, 91*(2), 295–304.

Beer, R. D. (1995). On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior, 3*(4), 471–511.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. J. Mataric, J. A. Meyer, J. Pollack, & S. W. Wilson (Eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 419–429). Cambridge: MIT.

Bekey, G. A. (2005). *Autonomous robots: From biological inspiration to implementation and control*. Cambridge: MIT.

Berg, H. C. (2004). *E. coli in motion*. New York: Springer.

Bernard, C. (1966[1878–1879]). *Leçons sur les phénomènes de la vie communs aux animaux et aux vegetaux*. Paris: Vrin.

Bernard, C. (1974). *Lectures on the phenomena of life common to animals and plants*. Springfield: Charles C. Thomas.

Braitenberg, V. (1984). *Vehicles. Experiments in synthetic psychology*. Cambridge: MIT.

Cannon, W. (1929). Organization for physiological homeostasis. *Physiological Reviews, 9*, 399–431.

Cannon, W. (1939[1932]). *The wisdom of the body* (2nd ed.). New York: W.W. Norton.

Capehart, B. L., & Terry, R. (1968). Digital simulation of homeostat modified to show memory and learning. *IEEE Transactions on Systems Science and Cybernetics, SSC4*(3), 188.

Descartes, R. (1988). *The passions of the soul* (Vol. 1, pp. 325–404). Cambridge: Cambridge University Press.

Di Paolo, E. (2000). Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. In J. A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat, & S. W. Wilson (Eds.), *From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behavior* (pp. 440–449). Cambridge: MIT.

Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429–452.

Di Paolo, E., Rohde, M., & Jaegher, H. D. (2010). Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart, O. Gapenne, & E. Di Paolo (Eds.), *Enaction: Towards a new paradigm for cognitive science*. Cambridge: MIT.

Egbert, M. D., Barandiaran, X. E., & Di Paolo, E. A. (2010). A minimal model of metabolism-based chemotaxis. *PLoS Computational Biology, 6*(12), e1001004.

Franchi, S. (2011a). Jammed machines and contingently fit animals: Psychoanalysis's biological paradox. *French Literature Series, 38*, 213–256.

Franchi, S. (2011b). Life, death, and resurrection of the homeostat. In S. Franchi & F. Bianchini (Eds.), *The search for a theory of cognition: Early mechanisms and new ideas* (pp. 3–51) Amsterdam: Rodopi.

Franchi, S. (2013). Homeostats for the 21st century? Lessons learned from simulating Ashby simulating the brain. *Constructivist Foundations, 8*(3), 501–532, with open peer commentaries and author's response

Haldane, J. S. (1917). *Organism and environment as illustrated by the physiology of breathing*. New Haven: Yale University Press.

Haldane, J. S. (1922). *Respiration*. New Haven: Yale University Press.

Haroules, G. G., & Haire, P. F. (1960). *Jenny: An improved homeostat* (Tech. Rep. AFCRC-TN-60-379), Air Force Cambridge Research Center.

Henderson, L. J. (1928). *Blood: A study in general physiology*. New Haven: Yale University Press.

Husbands, P., & Holland, O. (2008). The ratio club: A hub of british cybernetics. In P. Husbands, O. Holland, & M. Wheeler (Eds.), *The mechanical mind in history* (pp. 91–148). Cambridge: MIT.

Ikegami, T., & Suzuki, K. (2008). From a homeostatic to a homeodynamic self. *Biosystems, 91*(2), 388–400.

Izquierdo, E. J., Aguilera, M., & Beer, R. D. (2013). Analysis of ultrastability in small dynamical recurrent neural networks. In *Advances in Artificial Life, ECAL 2013. Proceedings of the Twelfth European Conference on the Synthesis and Simulation of Living Systems* (pp. 51–58). Cambridge: MIT.

Jonas, H. (1966). *The phenomenon of life: Towards a philosophical biology*. New York: Harper and Row.

Jonas, H. (1984). *The imperative of responsibility. Foundations of an ethics for the technological age*. Chicago: The University of Chicago Press.

Jonas, H. (1996). Evolution and freedom: On the continuity among life-forms. In *Mortality and morality. A search for the good after Auschwitz* (pp. 59–74). Evanston: Northwestern University Press.

Kant, I. (1952). *The critique of judgment*. Oxford: Oxford University Press.

Kant, I. (1993). *Opus postumum*. Cambridge: Cambridge University Press.

Kant, I. (1999). *Practical philosophy*. Cambridge: Cambridge University Press.

Malgrem, H. (2013). From Fechner, via Freud and Pavlov, to Ashby. *Constructivist Foundations, 9*(1), 104–105.

Michel, O. (2004). Webots: Professional mobile robot simulation. *Journal of Advanced Robotics Systems, 1*(1), 39–42.

Pfeifer, R., & Bongard, J. (2007). *How the body shapes the way we think: A new view of intelligence* (A Bradford book). Cambridge: MIT.

Pickering, A. (2010). *The cybernetic brain: Sketches of another future*. Chicago: The University of Chicago Press.

Plato. (1997). *Complete works*. Indianapolis: Hackett.

Starling, E. H. (1923). The wisdom of the body. The Harveian oration, delivered before the royal college of physicians of London on St. Luke's day, 1923. *British Medical Journal, 2*(3272), 685–690.

Varela, F. J. (1979). *Principles of biological autonomy*. New York: North-Holland.

von Foerster, H., Mead, M., & Teuber, H. L. (Eds.). (1953). *Cybernetics. Circular Causal and Feedback Mechanisms in Biological and Social Systems, Josiah Macy, Jr. Foundation, New York, transactions of the Ninth Conference*, New York, March 20–21, 1952.

Walter, W. G. (1961). *The living brain*. Harmondsworth: Penguin Books.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences, 1*(2), 97–125.

Wilkins, M. G. (1968). A new homeostat (Tech. Rep. 8.3, BCL), Biological Computer Laboratory, University of Illinois at Urbana-Champaign.

Chapter 18
Ad Hoc Hypotheses and the Monsters Within

Ioannis Votsis

Abstract Science is increasingly becoming automated. Tasks yet to be fully automated include the conjecturing, modifying, extending and testing of hypotheses. At present scientists have an array of methods to help them carry out those tasks. These range from the well-articulated, formal and unexceptional rules to the semi-articulated and variously understood rules-of-thumb and intuitive hunches. If we are to hand over at least some of the aforementioned tasks to machines, we need to clarify, refine and make formal, not to mention computable, even the more obscure of the methods scientists successfully employ in their inquiries. The focus of this essay is one such less-than-transparent methodological rule. I am here referring to the rule that ad hoc hypotheses ought to be spurned. This essay begins with a brief examination of some notable conceptions of ad hoc-ness in the philosophical literature. It is pointed out that there is a general problem afflicting most such conceptions, namely the intuitive judgments that are supposed to motivate them are not universally shared. Instead of getting bogged down in what ad hoc-ness exactly means, I shift the focus of the analysis to one undesirable feature often present in alleged cases of ad hoc-ness. I call this feature the 'monstrousness' of a hypothesis. A fully articulated formal account of this feature is presented by specifying what it is about the internal constitution of a hypothesis that makes it monstrous. Using this account, a monstrousness measure is then proposed and somewhat sketchily compared with the minimum description length approach.

Keywords Ad hoc • Scientific methodology • Minimum description length • Philosophy of artificial intelligence • Computational science

I. Votsis (✉)
New College of the Humanities, London
e-mail: ioannis.votsis@nchlondon.ac.uk

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_18

18.1 Introduction

Science is increasingly becoming automated. Tasks yet to be fully automated include the conjecturing, modifying, extending and testing of hypotheses. At present scientists have an array of methods to help them carry out those tasks. These range from the well-articulated, formal and unexceptional rules to the semi-articulated and variously understood rules-of-thumb and intuitive hunches. If we are to hand over at least some of the aforementioned tasks to machines, we need to clarify, refine and make formal, not to mention computable, even the more obscure of the methods scientists successfully employ with some measure of success in their inquiries. The focus of this essay is one such less-than-transparent methodological maxim for which much confusion and disagreement persists. I am here referring to the maxim that ad hoc hypotheses ought to be spurned. The need to throw light on this maxim and, in particular, on the notion of ad hoc-ness becomes all the more obvious when one considers that it is routinely invoked by scientists in assessing modifications or extensions of existing hypotheses and conjectures of new ones.

This essay begins with a brief examination of some notable contributions to the philosophical literature on ad hoc-ness, focusing, in particular, on two conceptions. After raising some problems specific to these conceptions, it is pointed out that there is a more general problem, namely the intuitive judgments that are supposed to motivate such conceptions are not always shared and indeed are sometimes even in conflict. The concept of ad hoc-ness is already burdened with too much intuitive baggage, signifying too many different things to too many different people. Instead of getting bogged down in what ad hoc-ness exactly means, I shift the focus of the analysis to one undesirable feature that is often, but not universally, present in alleged cases of ad hoc-ness. I call this feature the 'monstrousness' of a hypothesis for reasons that will become clear below. A fully articulated formal account of this feature is given by specifying what it is about the internal constitution of the content of a hypothesis that makes it monstrous. Using this account, a monstrousness measure is then proposed and somewhat sketchily compared with the minimum description length approach to simplicity/ad hoc-ness that is so popular in information and computer learning theory. It is argued that the monstrousness measure has some definite advantages over the minimum description length approach, at least with respect to one way the latter is conceived, but also some disadvantages. The main such disadvantage is that, as it stands, the measure is impracticable. The essay concludes with a proposal that seeks to extract a more practicable version of the measure. In doing so, the hope is that this essay helps prepare the ground for the delegation of a full gamut of scientific duties to the machines of the future.

18.2 Ad Hoc-Ness

There is a hefty amount of confusion surrounding the notion of ad hoc-ness. Given the notion's prevalence in everyday discourse, a dictionary entry makes for an apt starting point to our investigation. According to the Oxford English Dictionary, 'ad hoc' means "formed, arranged, or done for a particular purpose only".[1] This ordinary conception of ad hoc-ness is reflected in compound expressions like 'ad hoc committee'. But what does it mean for a hypothesis to be ad hoc? By and large, philosophers of science and other interested parties to this debate find the ordinary conception of ad hoc-ness deficient as an answer to this question. That's where their agreement ends, however, as they quarrel over how best to go about answering it. The result is that several conceptions of ad hoc-ness have arisen through the years.

Although the philosophical literature on ad hoc-ness is far from vast, there is more than enough material to prevent one from doing it justice in a short essay like this. I will thus restrict my comments to some notable contributions. Aside from Popper (1972) and Zahar (1973), both of which will be discussed in a bit more detail below, I would like to mention two other contributions in passing. The first is Leplin (1975). He identifies no less than five individually necessary and jointly sufficient conditions for ad hoc-ness, namely experimental anomaly, justification, tentativeness, consistency and non-fundamentality. Alas, his view suffers from a number of serious problems (see, for example, Grünbaum 1976) and is ultimately premised on the misguided idea that there is a unique "concept of '*ad hoc*ness' which the scientific community employs" (p. 316) [original emphasis]. Why it is misguided will, I hope, become clear in the next section. The second contribution is Forster and Sober (1994). Here it is contended that the term 'ad hoc' signifies unreasonable revisions to scientific theories and, in particular, to auxiliary hypotheses. Though Forster and Sober do not attempt to pin down the notion, they allude to an important connection between non-ad hoc-ness and simplicity.[2] The suggestion, roughly, is that simpler hypotheses are less ad hoc. We return to this connection in Sect. 18.5 below.

Let us now turn to Popper (1972). Popper's conception attempts to unpack the specificity of an ad hoc hypothesis in terms of its lack of excess testable content. In his own words:

> *Ad hoc* explanations are explanations which are not independently testable . . . In order that the *explicans should not be ad hoc, it must be rich in content: it must* have a variety of testable consequences, and among them, especially, testable consequences which are

[1] The same dictionary traces the Latin expression, which literally means 'for this', to the middle of the sixteenth century. See: http://www.oxforddictionaries.com/us/definition/american_english/ad-hoc?q=ad+hoc

[2] Simplicity, according to Forster and Sober, ought to be understood formally in terms of the Akaike information criterion. This criterion provides a method for selecting hypotheses by estimating their predictive accuracy. It does so by taking into account the trade-off between the simplicity of a hypothesis and its goodness-of-fit toward the data. For more details, see Akaike (1974).

> different from the *explicandum. It is these different* testable consequences which I have in mind when I speak of *independent tests, or* of *independent evidence* (1972, pp. 15–16, 193) [original emphasis].

Its connection to the ordinary conception of ad hoc-ness should be obvious. If a hypothesis H, which in this context is taken to be the explicans, has no excess testable content over explicandum E, then its purpose seems at best restricted to that of attempting to explain E. If, however, it has excess testable content, then the hypothesis has a broader purpose in that it can potentially explain other evidence that may turn out to be true. Following Popper, we may illustrate this conception with a non-ad hoc example from the history of science. The hypothesis positing the existence as well as the orbital and mass characteristics of the planet Neptune helps explain the perturbed orbit of Uranus within the Newtonian paradigm. That same hypothesis, however, has excess testable content over and above the perturbed orbit of Uranus. Among other things, it predicts (indeed to some extent successfully) additional perturbations in the orbits of other planets as well as that of the Sun around the solar system's barycenter.

Various problems afflict Popper's conception, one of which will be mentioned here. The problem I have in mind questions the idea that excess testable content is a sufficient condition for non-ad hoc-ness. Take *any* explicans that we would all, or at least the Popperians, judge to be ad hoc. We can easily turn it into one that Popperians would deem non-ad hoc simply by conjoining to it *any* random proposition whose testable content exceeds that of the explicandum. Here's an example. Suppose that the following hypothesis Z_1 offers an ad hoc explanation of S_1.

Z_1: Zeus exists and he is sometimes angry and whenever he is angry he lights up the sky with thunderbolts.
S_1: Sometimes the sky lights up with thunderbolts.

Suppose, moreover, that we add a random proposition A_1 to the explicans. It doesn't matter whether this proposition is true or false – in this case we happen to choose a true one.

A_1: Free falling objects near the earth's surface accelerate roughly at 9.81 m/s^2.

Following Popper's conception, conjunction $Z_1 \wedge A_1$ is not an ad hoc explanation of S_1 for it has excess testable (and in fact tested) content, namely that concerning the rate of acceleration of freely falling objects near the surface of the earth. But, surely, the conjunction is no less ad hoc than before. In other words, having excess testable content does not guarantee non-ad hoc-ness. At best, having excess testable content is a necessary condition for non-ad hoc-ness.[3]

[3]Grünbaum (1976, p. 343) notes Popper's ambiguous behaviour towards the logical status of the condition of excess testable content, sometimes treating it as merely a sufficient condition and sometimes as both necessary and sufficient for non-ad hoc-ness.

Consider next Zahar's approach. I here focus on one of three proposed conceptions put forth by him.[4] A theory, holds Zahar, is ad hoc "if it has no novel consequences as compared with its predecessor" (1973, p. 101). A consequence or prediction is novel so long as the corresponding phenomenon was not the explanatory target of the scientists who constructed the theory. For example, the perihelion of the planet Mercury is a novel prediction of the general theory of relativity because the perihelion was presumably not explanatorily targeted by Einstein in his construction of the theory.[5] There are some differences between Zahar's and Popper's conceptions. For example, Zahar construes ad hoc-ness as a relation between successive hypotheses, not, as Popper does, simply as a kind of failed explanatory relation. Even so, ad hoc-ness judgments elicited from the two conceptions are, on the face of it at least, often in agreement. A hypothesis possessing consequences whose corresponding phenomena were not explanatorily targeted during construction (and where those phenomena were not novel for its predecessor) has excess testable content over and above the phenomena it was constructed to explain. That's not to say that judgments elicited from the two conceptions are never divergent however. A successor hypothesis with a solitary novel consequence has excess testable content and hence is non-ad hoc for Popper but counts as ad hoc for Zahar if that consequence is also novel for its predecessor.[6]

Zahar's conception fares no better than Popper's. Two problems stand out. First, just because a successor hypothesis makes no novel predictions compared to its predecessor does not mean that it is any worse off or indeed ad hoc. Some progress in science involves removing ad hoc elements from a predecessor hypothesis to obtain a non-ad hoc, or less as hoc, successor hypothesis. Ridding an otherwise empirical hypothesis from supernatural posits is exactly one such type of progress. Second, suppose that a predecessor hypothesis H explains all and only O_1. Suppose moreover that two scientists, F and G, independently construct the same successor hypothesis H', which explains all and only O_1 and O_2. Suppose, finally, that O_2

[4]Here are the other two: "It is *ad hoc*$_2$ if none of its novel predictions have been actually 'verified' . . . [a] theory is said to be *ad hoc*$_3$ if it is obtained from its predecessor through a modification of the auxiliary hypotheses which does not accord with the spirit of the heuristic of the programme" (1973, p. 101) [original emphasis]. Grünbaum (1976, p. 341) notes the similarities between Zahar's notion *ad hoc*$_1$ and his own notion ad hoc (c). For a more detailed discussion of Zahar's notions, particularly *ad hoc*$_2$, the reader may consult Redhead (1978).

[5]This example has been contested by Earman and Glymour (1978) who argue, convincingly, that the perihelion was in fact explanatorily targeted. Less controversial examples include all the cases involving temporally novel phenomena, i.e. phenomena which were not known at the time a hypothesis was constructed and hence could not have been explanatorily targeted by the constructors.

[6]How much divergence exists between judgments elicited from the two conceptions depends on a number of factors. For example, if Zahar permits the notion of novel consequences to also range over consequences that cannot be tested, then the divergence is significant.

was explanatorily targeted by G but not by F. Eliciting a judgment from Zahar's conception of ad hoc-ness yields a contradiction: H' turns out to be both ad hoc and non-ad hoc.[7]

18.3 Taking a Step Back

Is it not immature to stop our investigation into conceptions of ad hoc-ness after considering only two of them? Yes, but there is good reason to do so. The existing conceptions, as well as at least some of the objections levelled against them, rely heavily on intuitive judgments about what are genuinely ad hoc or non-ad hoc cases and what are essential and what accidental features of ad hoc-ness. The problem is that these intuitive judgments are not always shared. In fact, such judgements are sometimes in conflict. As an illustration of this conflict, take the intuition that ad hoc-ness is always undesirable. First off, note that this intuition is not inherited from the ordinary conception of ad hoc-ness. Ad hoc committees, for example, serve a more limited role than non-ad hoc ones but that doesn't make them any less desirable. Secondly, and more importantly, it is highly doubtful that there is a uniform meaning of ad hoc-ness in the non-ordinary, i.e. scientific, context. Forster and Sober, for instance, are categorical in their condemnation of ad hoc hypotheses: "we reserve the term 'ad hoc' for revisions of the bad kind" (1994, p. 17). But not everybody agrees. Grünbaum, for example, holds that "I evidently do not deny that *certain* uses of the term 'ad hoc' are intended to be derogatory" (1976, p. 361) [first emphasis added]. And Popper is of two minds. He sometimes claims that "the corresponding requirement that explanations of this kind [i.e. ad hoc] should be avoided [is], I believe, among the main motive forces of the development of science" (1972, p. 192). But at other times he suggests that "... we must not exclude all immunizations, not even all which introduce ad hoc auxiliary hypotheses" (1974, p. 32). More generally, and perhaps more relevantly, Holton (1969) puts paid to the idea that scientists themselves share a unique conception of what it means for a hypothesis to be ad hoc:

> The scientist who adopts somebody's hypothesis or creates his own for a specific purpose, "in order to account" for a bothersome result or feature of the theory, regards it as ad hoc - not necessarily in a derogatory sense... Thus we have found in the scientific literature characterizations of the following kinds for acceptable ad hoc hypotheses: "not inconceivable," "reasonable," "plausible," "fundamental," "natural," "appealing," "elegant," "likely," "assumed a priori to get the desired results," "auxiliary" or "working hypothesis." On the other hand, when an ad hoc hypothesis is rejected, we see it described in the following

[7] Strictly speaking, H' is ad hoc for G but non-ad hoc for F. If Zahar were a subjectivist then he could perhaps get away with this reply by claiming that ad hoc-ness is a subjective matter. The trouble is he is not – see his comments (1973, pp. 103–104). What is more, if ad hoc-ness is to have any epistemic import it could not be something that varies from subject to subject. For more on this and related problems see Votsis (2014).

way: "artificial," "complex," "contrived," "implausible," "bothersome," "unreasonable," "improbable," "unlikely," "unnecessary," "ugly." (p. 178)

How do we proceed in light of such discord? Instead of attempting to disentangle the intuitions behind the use of the term 'ad hoc-ness', I propose that we focus on one feature – or, otherwise put, one consistent set of intuitions – that is often associated with alleged cases of ad hoc-ness.[8] This is a feature that I deem to be undesirable at all times for a hypothesis to possess. For this reason it is useful that it be given a clear conception. After all, in matters of sound methodology and epistemology we need concepts and rules that tell us in as unambiguous a way as possible what is a legitimate and what an illegitimate modification of a central hypothesis and/or an auxiliary, which hypotheses are likely to be true and which false, what evidence weighs more and what less, etc. In order to avoid any carryover from the intuitive baggage associated with the term 'ad hoc', or the intuitive baggage of any other closely related term for that matter, I suggest that we use a relatively unsullied term to express this feature. For reasons that will soon become apparent, I propose the use of the term 'monstrousness'.[9]

18.4 Disjointedness and Monstrousness

The characteristic I have in mind is something that has not gone unnoticed in discussions of ad hoc-ness. Grünbaum, for example, asserts that what is undesirable about some ad hoc hypotheses is that they are "artificial, contrived or arbitrary" (1976, p. 358). Instead of trying to expand on what any of these terms truly mean, a move which will inevitably pull us back down into controversy, I will instead employ the largely untainted term 'monstrousness'. The reason why we call this characteristic thus is that it indicates the extent to which a hypothesis is assembled out of confirmationally disjointed content parts, in a manner similar to the way the most famous monster in literature, i.e. the monster in Mary Shelley's Frankenstein, is assembled out of a motley of parts.[10] Numerous examples of undesirable ad hoc hypotheses exhibit this disjointedness in good measure. This is certainly true of $Z_1 \wedge A_1$. It is, however, less pervasive and perhaps even absent in hypotheses that are typically more desirable, e.g. the conjunction of the Newtonian paradigm with the hypothesis that Neptune exists and has certain orbital and other characteristics.

How are we meant to understand (confirmational) disjointedness?[11] I propose the following articulation:

[8] One possible approach which is eschewed here, though it is legitimate, is to admit several distinct notions of ad hoc-ness.

[9] A version of this notion was first explored in Votsis (2014).

[10] For a discussion of the notion of content part you may consult Gemes (1994, 1997).

[11] Elsewhere, see Votsis (2015), I call this notion 'confirmational disconnectedness'. Either name is fine.

Disjointedness: Any two content parts of a non-self-contradictory proposition Γ expressed as propositions A, B are disjointed if and only if for all pairs of internally non-superfluous and non-sub-atomic propositions α, β where α is a relevant deductive consequence of A and β is a relevant deductive consequence of B: (i) there is no true or partly true proposition γ such that γ is a relevant deductive consequence of α and γ is also a relevant deductive consequence of β, (ii) $P\ (\alpha/\beta) = P(\alpha)$ where $0 < \mathrm{P}(\alpha), \mathrm{P}(\beta) < 1$ and (iii) there is no atomic proposition δ that is a relevant deductive consequence of $\alpha \wedge \beta$ and is not a relevant deductive consequence of either α or β on their own.

Hence jointedness can be articulated thus:

Jointedness: Any two content parts of a non-self-contradictory proposition Γ expressed as propositions A, B are jointed if and only if for some pair of internally non-superfluous and non-sub-atomic propositions α, β where α is a relevant deductive consequence of A and β is a relevant deductive consequence of B: either (1) there is a true or partly true proposition γ such that γ is a relevant deductive consequence of α and γ is also a relevant deductive consequence of β, or (2) $P(\alpha/\beta) \neq P(\alpha)$ where $0 < \mathrm{P}(\alpha), \mathrm{P}(\beta) < 1$ or (3) there is at least one atomic proposition δ that is a relevant deductive consequence of $\alpha \wedge \beta$ but is not a relevant deductive consequence of either α or β on their own.

Let us take a closer look at the notion of disjointedness. What is a content part? Roughly put, a content part c of Γ is a non-trivial consequence of Γ from which one cannot derive Γ itself. In other words, c is strictly smaller in content than Γ. The non-superfluousness clause is there to remove superfluous content from the evaluation. One reason for doing so is that it reduces the evaluation's complexity. Note that the clause wards off superfluousness only within each of α or β, but not across them, i.e. between α and β. In short, the contents of α and β are allowed to overlap. The reason we permit such overlap is that it allows the detection of one kind of jointedness.

Now consider clause (ii). The probabilities involved are meant to be objective. That is, they are meant to indicate true relative frequencies and/or true propensities of events, states-of-affairs, properties, etc., expressed by propositions. This is an important qualification as it de-subjectivises the notion of disjointedness (and hence jointedness) – we briefly return to this issue in the next section. The notion of probabilistic independence allows us to make an important first step in expressing the idea that two content parts are confirmationally disjointed. For if α, β are probabilistically independent we know that the probability of the one is not affected if we assume something about the truth (/falsity) of the other. This connects well with the relevance criterion of confirmation various Bayesians live by according to which e stands in a confirmational relation to (i.e. either confirms or disconfirms) h if and only if the two are probabilistically dependent.

Properly accounting for the confirmational disjointedness of two propositions A, B requires that we inspect not only their total content but also the content of their parts. That's because the two propositions may be probabilistically independent even though some of their parts are not. That's where the notion of *deductive consequence* comes in handy. By checking whether each and every – see the next paragraph for a qualification – deductive consequence of the one proposition is probabilistically independent from/dependent on each and every deductive consequence of the other we ensure that we take all confirmation relations between

A, B into account. Consider the following example as an illustration. Suppose that A: $U_1 \wedge U_2$ and B: $U_1 \wedge V_1$, that $P(U_1/U_2) = P(U_1)$ and $P(U_1/V_1) = P(U_1)$ and that $P(U_1) = 0.5$, $P(U_2) = 0.5$, $P(V_1) = 0.5$ and $P(B/A) = 0.25$. From the fact that U_1 and U_2 are probabilistically independent we can derive that $P(U_1 \wedge U_2) = P(U_1) * P(U_2) = 0.25$. So we know that $P(A) = 0.25$. Similarly from the fact that U_1 and V_1 are probabilistically independent we can derive that $P(U_1 \wedge V_1) = P(U_1) * P(V_1) = 0.25$ and thus we know that $P(B) = 0.25$. Using Bayes theorem we can thus derive that $P(A/B) = 0.25$. But that just means that A, B are probabilistically independent since $P(A/B) = P(A)$. But notice that the same is not true of all the consequences of A, B. Take U_1. It is a consequence of A and of B. But $P(U_1/U_1) = 1$ and hence $P(U_1/U_1) \neq P(U_1)$. Thus, there is a proposition α and a proposition β, namely U_1 in both cases, such that $P(\alpha/\beta) \neq P(\alpha)$ and hence A, B turn out to be jointed when we take a closer look. To recap, in order to make sure that two propositions are confirmationally disjointed we must demand that probabilistic independence holds *all the way down*.

As already alluded, we do not actually care about *all* deductive consequences. This is because some of them are trivial. In fact, were we to take these into account we would render the concept of disjointedness unsatisfiable. This can be demonstrated with a simple example. Regardless of the exact content of A, B there are always validly derivable but trivial consequences of each that they have in common, e.g. $A \vee B$. Such trivial common consequences guarantee the existence of a pair of propositions α_i, β_i for which $P(\alpha_i/\beta_i) \neq P(\alpha_i)$ provided $0 < P(\alpha_i) < 1$. Otherwise put, it guarantees that A, B are not disjointed. To rule out such cases we restrict our attention to all *relevant* deductive consequences. The notion of relevance can be found in Schurz (1991) where he explains that "the conclusion of a given deduction is irrelevant iff the conclusion contains a component [i.e. a formula] which may be replaced by any other formula, salva validitate of the deduction" (pp. 400–401).[12]

Now consider clause (i). If there were such a proposition γ, α and β would share content and hence any confirmation of that content would confirm at least a content part of each α, β. In short, such a proposition γ would force us to admit that α and β and hence A and B are confirmationally linked. That's why clause (i), *qua* a disjointedness clause, requires that there is no such proposition γ.

Note that although two propositions A, B may be probabilistically independent all the way down, they may still be confirmationally related through jointly and relevantly – as in 'having a relevant deductive consequence' – entailing a proposition δ that neither entails on its own and whose truth would confirm both.[13] Clause (iii)

[12] There is also an analogous notion that applies to predicates – see Schurz (2014).

[13] An example can be garnered from discussions of causal modelling. Two causes may be probabilistically independent and yet their presence may be sufficient to yield a joint effect. The presence of the joint effect confirms the presence of both causes. Thus, the two causes are confirmationally related even though they and the propositions expressing their presence are probabilistically independent – the latter presumably all the way down. Another example may be sourced from the domain of mathematics. Axioms are probabilistically independent (presumably

is there to ensure that there is no 'indirect' confirmational relation between A and B via such a consequence δ. If there is such a consequence then A, B are not disjointed. Not just any joint consequence will do. Unless we place some restrictions on what counts as a proposition δ, the concept of disjointedness is once again rendered unsatisfiable. That is, if we required only that there is no δ that any α, β jointly (but not individually) entail then there would almost always be such a δ, more or less notwithstanding what content A, B possess. For example, one such δ is $\alpha \wedge \beta$ where α is not logically equivalent to β . More generally, a proposition ζ that is jointly (but not individually) entailed by two propositions α, β cannot play the role of δ if it is logically equivalent to $\omega \wedge \varepsilon$ where both ω is a relevant consequence of α and ε is a relevant consequence of β. Such joint consequences are trivial for our purposes and therefore incapable of assisting is us in our quest to find a confirmational relatedness between A, B that is not captured by conditions (i) and (ii). Note that to dismiss such trivial consequences we cannot rely on the simple notion of relevant consequence like we did before since $A \wedge B$ is a relevant deductive consequence of $A \wedge B$, i.e. it is not the case that we can substitute any formula in it with any other formula without disturbing the validity of the deduction. Instead, to solve our problem we must rely on the notion of an atomic proposition. This not the same as an atomic sentence or proposition in the logical sense of the word. Rather, in the current sense (and roughly speaking) an atomic proposition contains the least amount of content that does empirically significant work. More formally, a proposition φ is atomic if, and only if, φ is non-superfluous and truthfully represents all and only the content of an atomic state of the world. Crucially, the content of an atomic proposition cannot be decomposed into distinct atomic or molecular content parts. How does it rule out guaranteed joint consequences like $\alpha \wedge \beta$? If δ is atomic and relevantly follows from $\alpha \wedge \beta$, then its potential decomposition into an equivalent conjunction could not be the conjunction $\alpha \wedge \beta$. For suppose δ is equivalent to $\alpha \wedge \beta$. Recall that α, β are (by stipulation) not sub-atomic. That means that either one or both are atomic or molecular in which case δ is not atomic. Contradiction! In other words, requiring atomicity forbids such decompositions and hence consequences like $\alpha \wedge \beta$ are ruled out.

What use could we possible have for disjointedness? Well, disjointedness forms a barrier against the spread of confirmation. Thus even though monstrous hypotheses get confirmed under this view, the confirmation they receive for a content part that is disjointed from other content parts doesn't spread to those other parts. For example, the truth of S_1 confirms (a part of) Z_1 precisely because Z_1 was designed to entail S_1 but, crucially, this confirmation does not spread to the non-S_1 part of Z_1, namely the part that asserts that Zeus exists and posits the existence of Zeus and the property that he is sometimes angry. The approach just outlined is similar to Schurz (2014) in that it aims to regulate how confirmation spreads within the content parts of a hypothesis. Unlike him, however, I insist that we are still dealing with a case of

all the way down) but two or more of them may be necessary to derive a single theorem. This last example is only meant as a crutch to help understand condition (ii). The view I am proposing here is restricted to empirical, not mathematical, hypotheses.

genuine confirmation when the support gained from some piece of evidence does not spread to content parts other than those corresponding to the evidence. My approach thus offers a unified treatment of confirmation relations.

Hypotheses may possess both disjointed and non-disjointed content parts. To be exact, since disjointedness and non-disjointedness are relations that hold between various content parts of hypotheses, the claim is that hypotheses may possess content parts, some of which are disjointed and others non-disjointed to other content parts. It is my conjecture that almost all hypotheses will have some disjointed parts and hence will be monstrous to some extent. It thus makes sense to devise a way to measure the level of monstrousness of a hypothesis. But before we do that we must consider one last complication.

Should judgments of monstrousness be affected by the way the same content is distributed between two different propositions? We are obviously free to cut up content any which way we like, i.e. content distribution is an arbitrary affair. Being arbitrary means that it doesn't tell us anything about the world. Hence, in answer to the above question, the way in which content is distributed should not affect our monstrousness judgements. Take two propositions A_1, B_1 where A_1: $D_1 \wedge D_2 \wedge D_3$ and B_1: E_1. Suppose that when we compare these propositions using some measure of monstrousness we get some score w_1. Now take another two propositions A_1', B_1' where A_1': $D_1 \wedge D_2$ and B_1': $E_1 \wedge D_3$. Suppose that when we compare A_1', B_1' using the same measure of monstrousness the resulting score is w_2. According to the above argument, since $A_1 \wedge B_1$ has the same content as $A_1' \wedge B_1'$ any proposed measure of monstrousness should ensure that $w_1 = w_2$. That is to say, monstrousness judgments should be invariant under different ways of distributing the same content between two different propositions.

One, perhaps the only, way to pull this off is to calculate monstrousness scores on the basis of *all* distinct ways of distributing the same content between two different propositions. In what follows, I put forth a proposal of exactly such a measure. The proposed measure is not the final word on the matter but still worth considering since, in my view, it is heading in the right direction.[14] Without further ado, here's the proposal: The monstrousness of a proposition is given by the ratio of the sum of disjointed pairs of parts taken from all distinct ways of distributing its content to the sum of the total number of pairs of parts (i.e. jointed and disjointed) taken from all distinct ways of distributing its content. Formally, the monstrousness m of a proposition Δ is given by the following function:

$$m(\Delta) = \sum_{i=1}^{n} d_i^{\alpha,\beta} / \sum_{i=1}^{n} t_i^{\alpha,\beta}$$

[14]One reason for its inadequacy is that when a pair of propositions α, β are deemed jointed the strength of their jointedness, e.g. the degrees of their probabilistic dependence, is neglected. Arguably such information should have a role in any suitably sensitive measure of monstrousness.

where $d_i^{\alpha,\beta}$ denotes the number of disjointed pairs α, β in a given content distribution i, $t_i^{\alpha,\beta}$ denotes the total number of jointed plus disjointed pairs α, β in a given distribution i, and n denotes the total number of content distributions.[15]

The number of disjointed pairs in a given content distribution is determined by counting how many times a different pair of relevant deductive consequences α, β turns out to satisfy clauses (i) - (iii) at the same time. Any pair that is not disjointed is counted as jointed. The higher (/lower) the value of m the more (/less) monstrous the content expressed by Δ. Note that this value is the same no matter how we cut Δ since we take into account *all* other ways the same content can be distributed between two propositions.

One of the advantages of the proposed measure is that it is quite broad in its range of application. This is due to the fact that the notions of jointedness and disjointedness don't place any restrictions on the propositions being compared other than the restriction that they are consistent. As a result, the said propositions can be drawn from a large pool of entries which includes central hypotheses, auxiliaries, explanantia and explananda. This not only allows us to gauge the monstrousness of the most commonly touted relations, e.g. the relation between a central hypothesis and an auxiliary hypothesis or the relation between an explanans and an explanandum, but also of any other relation we can think of, e.g. the relation between one auxiliary hypothesis and another.

18.5 Monstrousness and MDL

This essay is meant to give some guidance, however limited, on the subject of how machines may automate the task of discriminating between bad and good hypotheses. With this aim in mind it is worth comparing, albeit briefly and superficially, my approach to a leading approach employed in information and computer learning theory, namely the minimum description length (MDL) approach. MDL is a hypothesis selection principle. The preferred hypothesis, according to this approach, is the one that provides the most economical description of the data. It is not hard to see how MDL is related to the demand for simpler hypotheses. Thus, Rissanen, MDL's founding father, doesn't hesitate to assert that "the notion of simplicity is entirely in line with the modern notions of complexity of description" (1983, p. 421). The notion of ad hoc-ness is not as prominent in this literature. But when it does make its appearance the central idea seems to be that simpler hypotheses are less ad hoc – an idea that, as we earlier saw, Forster and Sober also find attractive.

[15]This function only makes sense if the total number of content distributions is finite. I am assuming this is the case. Arguments for this assumption can be given but require quite a bit of stage-setting. For now, it suffices to say that this assumption is guaranteed to hold if the hypotheses in question can be fully decomposed into a finite number of atomic propositions.

> ... the *minimum description length principle* ... extends Occam's Razor to say that the best hypothesis is the one that minimizes the total length of the hypothesis plus the description of the *exceptions* to the hypothesis. The intuition is that ad hoc hypotheses are simply lists of examples, which makes them no shorter than the examples they purport to summarise. In contrast, good hypotheses reduce many examples to a simple, general rule (Shavlik and Dietterich 1990, p. 47) [original emphasis].

There are obvious similarities between my measure of monstrousness and MDL measures of simplicity. For example, a hypothesis that conjoins propositions that express disparate facts, e.g. a_1 is a white swan $\wedge$ a_2 is a white dwarf, gets a high monstrousness score *and* a low MDL simplicity score – in the latter case since its length is presumably no shorter than the combined length of the individual propositions it seeks to summarise.[16] But there are also differences. A hypothesis that conjoins propositions that express related facts, e.g. b_1 is a white swan $\wedge$ b_2 is a white swan $\wedge \ldots \wedge$ b_n is a white swan, but does not convey them in terms of a generalisation, e.g. all observed swans are white, gets a low simplicity score, at least on a 'face value' reading of the MDL account, for the abovementioned reason but *does not* get a high monstrousness score. The reason for the latter is that the individual facts are systematically related via common relevant deductive consequences, e.g. the claim that there are at least two white swans. A corollary is that under the monstrousness approach, and contra (a 'face value' reading of) MDL, a conjunction of propositions expressing related facts gets the same score as a generalisation of them since they both possess the same content, provided the clause 'and nothing else satisfies the antecedent' is added into the mix, and hence share the same distinct ways of distributing that content between two propositions.

Two further differences between these approaches, this time at a more abstract level, are worth pondering over. Monstrousness is determined by objective facts about the true relative frequencies and/or true propensities of events, states-of-affairs, properties, etc., expressed by the propositions being compared. Its determination is thus a thoroughly *a posteriori* matter and herein lies the strength of this approach. By contrast, the MDL approach to simplicity seems to reward or penalise hypotheses in a strongly *a priori* manner by insisting that the simplest hypothesis is to be preferred regardless of facts on the ground. That's the first difference between the two approaches. The second difference concerns practicality. Here the earlier mentioned strength of the monstrousness approach turns into a weakness. Since the probabilities involved are objective and we have limited access to these, the approach is severely handicapped in its practicability. The same is not true of the MDL approach, which, by virtue of its strong *a prioricity*, is capable of offering advice even in the face of limited access to facts on the ground.

Allow me to draw this section to a close by contemplating how one might go about turning the monstrousness approach into something more practicable. Recall

[16]There are also intimate connections between the minimum description length approach and Schurz's idea of a content element, since the latter is understood in terms of the *length* of a formula after it has been transformed to its negation-normal form.

that what this approach is supposed to measure is the confirmational (dis)jointedness between content parts of a hypothesis. Now, although we have limited access to the objective probabilities that determine the said confirmational (dis)jointedness we are not completely in the blind. One kind of information that is more readily accessible concerns the deductive relations between different content parts. For example, we can at least judge whether or not two content parts are *logically* independent. Moreover, we can judge whether or not any of their relevant (and non-redundancy containing) deductive consequences are logically independent. And we can even judge whether or not the conjunction of any pair of such consequences yields a relevant deductive atomic proposition. Although these judgments fall short of empowering a full assessment of the monstrousness of a hypothesis, they at least give us some hints about the general trajectory such an assessment ought to take. Whether or not these hints are more informative than MDL-derived recommendations remains to be seen. My hunch is that, allowing for suitable modifications, these two approaches should in principle be capable of reaching similar levels of informativeness.

18.6 Conclusion

To summarise: In the first part of the essay I briefly considered some prominent philosophical accounts of ad hoc-ness and argued that these are deficient in some important respects. I then made the case that appeal to intuitive judgments would only help adjudicate between rival conceptions of ad hoc-ness if those judgments were shared, something that is plainly not true. As an alternative, I recommended that we drop the intuitively-loaded term 'ad hoc-ness' and shift our focus onto a genuinely undesirable feature of hypotheses. I dubbed this feature that is often, but not universally, present in presumed examples of ad hoc-ness the 'monstrousness' of a hypothesis. I then proceeded to explicate this notion by means of the technical notions of probabilistic independence, content part, relevant consequence and atomic proposition. I followed that up with a proposed measure of the monstrousness of hypotheses. I subsequently, and admittedly fleetingly, compared the monstrousness measure with the general spirit of MDL approaches to simplicity/ad hoc-ness. The outcome of that comparison was that the former profits from its a-posteriori attitude towards the problem it studies but only at the expense of practicability while the latter does the opposite. The essay concluded with a suggestion of how to understand monstrousness in a more practicable manner, one that hopefully contains hints towards the potential automation of tasks that include the conjecturing, extending and modifying of hypotheses.

Acknowledgements My sincerest thanks to three anonymous referees as well as to my colleagues, Gerhard Schurz and Paul Thorn, for valuable feedback on the material presented in this essay. I acknowledge the German Research Foundation (Deutsche Forschungsgemeinschaft) for funding my research under project B4 of Collaborative Research Centre 991: The Structure of

Representations in Language, Cognition, and Science. Part of this essay has been written while working on the project 'Aspects and Prospects of Realism in the Philosophy of Science and Mathematics' (APRePoSMa) during a visiting fellowship at the University of Athens. The project and my visits are co-financed by the European Union (European Social Fund—ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF)—Research Funding Program: THALIS—UOA.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Earman, J., & Glymour, C. (1978). Einstein and Hilbert: Two months in the history of general relativity. *Archive for History of Exact Sciences, 19*, 291–308.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1–35.

Gemes, K. (1994). A new theory of content I: Basic content. *Journal of Philosophical Logic, 23*(6), 595–620.

Gemes, K. (1997). A new theory of content II: Model theory and some alternatives. *Journal of Philosophical Logic, 26*(4), 449–476.

Grünbaum, A. (1976). Ad hoc auxiliary hypotheses and falsificationism. *British Journal for the Philosophy of Science, 27*(4), 329–362.

Holton, G. (1969). Einstein, Michelson, and the "crucial" experiment. *Isis, 60*(2), 132–197.

Leplin, J. (1975). The concept of an ad hoc hypothesis. *Studies in History and Philosophy of Science, 5*(4), 309–345.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.

Popper, K. R. (1974). Replies to my critics. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 961–1197). La Salle: Open Court.

Redhead, M. (1978). Ad hocness and the appraisal of theories. *British Journal for the Philosophy of Science, 29*(4), 355–361.

Rissanen, J. (1983). Universal prior for integers and estimation by minimum description length. *The Annals of Statistics, 11*(2), 416–431.

Schurz, G. (1991). Relevant deduction: From solving paradoxes towards a general theory. *Erkenntnis, 35*, 391–437.

Schurz, G. (2014). Bayesian pseudo-confirmation, use-novelty, and genuine confirmation. *Studies in History and Philosophy of Science, 45*(1), 87–96.

Shavlik, J. W., & Dietterich, T. G. (Eds.). (1990). *Readings in machine learning*. San Mateo: Morgan Kaufmann Publishers.

Votsis, I. (2014). Objectivity in confirmation: Post hoc monsters and novel predictions. *Studies in History and Philosophy of Science, 45*(1), 70–78.

Votsis, I. (2015). Unification: Not just a thing of beauty. *Theoria: An International Journal for Theory, History and Foundations of Science, 30*(1), 97–114.

Zahar, E. (1973). Why did Einstein's programme supersede Lorentz's? (Part I). *British Journal for the Philosophy of Science, 24*, 95–123.

Chapter 19
Arguably Argumentative: A Formal Approach to the Argumentative Theory of Reason

Sjur K. Dyrkolbotn and Truls Pedersen

Abstract We propose a formal approach to the argumentative theory of reason, combining argumentation theory and modal logic in a novel way. We show that the resulting framework can be used to model important mechanisms identified by the theory, including how confirmation bias and other problematic modes of reasoning may in fact serve an important argumentative purpose that can give rise to classically sound conclusions through the process of social deliberation. We go on to suggest that the argumentative theory is based on an understanding of intelligent reasoning and rationality that sees these notions as irreducibly social, and that the argumentative theory itself provides a possible starting point in the search for new theoretical foundations based on this understanding. Moreover, we suggest that formal logic can aid in the investigation of foundational issues, and we sketch the development of an axiomatic approach to the study of rational deliberation.

Keywords Rationality • Argumentation • Deliberation • Cognitive bias • Modal logic

19.1 Introduction

The idea that social interaction and rationality are mutual dependent notions is gaining ground in many different fields of research, including economy, law, biology and artificial intelligence (Blume and Durlauf 2001; Terrell 2012; Waal and Ferrari 2010; van Benthem 2011; Ossowski 2013). In all these areas, there is a trend towards viewing rationality as fundamentally embedded in a social context, a context that is seen as important not only because people are social and tend to interact, but also because *who* they are, *how* they think, and *what* they want tends to depend on how

S.K. Dyrkolbotn (✉)
Department of Philosophy and Religious Studies, Utrecht University, Utrecht, The Netherlands
e-mail: s.k.dyrkolbotn@durham.ac.uk

T. Pedersen
Department of Information Science and Media Studies, University of Bergen, Bergen, Norway

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_19

they engage with each other and their environment.[1] With this as a starting point, intelligence can no longer be viewed merely as an ability to make optimal choices, it must involve also the ability to interact meaningfully with other agents, to negotiate meaning, reassess goals, and formulate new agendas. Intelligence, in particular, is embedded in a social discourse, and part of its function is reflexive; intelligent agents can come together to change those very parameters by which we (and they) assess what counts as "rational" in a given environment.[2]

To accommodate this point of view across different domains, we need better theoretical foundations, allowing us to investigate the relationship between reasoning and interaction, based on the starting point that they are co-dependent and co-evolving. In this paper, we argue that this challenge can be addressed using formal logic, drawing on tools and techniques developed in the context of multi-agent systems.[3] In particular, we propose a formal approach to the argumentative theory of reason, introduced in Mercier and Sperber (2011). The key idea is that reasoning evolved to facilitate efficient argumentation, not necessarily to help us arrive at logically correct forms of inference or in making reasonable decisions.

This can have important implications for logical modeling of rational interaction and in this paper we sketch the development of a general logical framework that enables us to capture the idea formally. Our overarching aim is to argue that a formal approach to the argumentative theory of reason can provide interesting new insights, particularly regarding the socially emergent nature of rationality.

The structure of the paper is as follows. In Sect. 19.2, we present the argumentative theory, focusing on its implications for our understanding of agency. We

[1] This perspective has long been influential in political philosophy, sociology and social psychology, particularly in research traditions going back to the work of George Herbert Mead and the Chicago school (Mead 1967). But to many other fields, particularly those based on formal methods or rational actor models, it represents an important recent trend, a move away from methodological individualism towards more holistic approaches. Important work has also been devoted to attempting to unite the two paradigms, such as List and Dryzek (2003) which presents a deliberative approach to social choice theory.

[2] This does not mean that the various intelligence-as-optimization theories that have been proposed (see e.g., Russell 1997) are mistaken. On the contrary, we agree that such theories can be highly informative. However, they are also inherently incomplete. Therefore, they should be complemented by models that allow us to investigate other aspects, such as reflexive reasoning about what it means for an agent to optimize its behavior. If we ever succeed in creating a truly intelligent agent, such an agent might well take issue with our explanation of what exactly it is that makes it intelligent. Indeed, such an ability would in itself be a mark of intelligence, perhaps the best we can hope for.

[3] The connection between various branches of social science and formal logic and computer science has received much attention in recent years and it has led to a surge of interest in interdisciplinary research (Parikh 2002; van Benthem 2008; Verbrugge 2009). However, while much recent work in applied logic has been devoted to modeling agency and interaction, the usual starting point is still that agents reason in adherence to some given standards of correctness, which remain fixed even in models that are designed specifically to model changes in the environment of the agents. In particular, most formal work is based on an individualistic and highly normative view on what it means to reason correctly about a state-of-affairs, a view we will challenge in this paper.

argue that the notion of argumentation that is at work challenges existing traditions in argumentation theory, particularly formal theories, and we go on to present an alternative formalization which looks at argumentative structures in a new light, as the basis upon which agents' subjective interpretations of the world are formed. In short, we introduce *argumentative agents*, and we argue that they should be studied further.

In Sect. 19.3, we go on to present a logical formalism for studying what we call *argumentative deliberation*, interaction between argumentative agents that can serve to generate novel interpretations of the world. Since our purpose in this paper is to focus on main ideas rather than technical details, we present a simple logical framework, containing some basic constructions which can be developed further using existing tools from modal logic. We show through examples that this is sufficient to allow us to capture essential aspects of the mechanisms addressed in Mercier and Sperber (2011), and we sketch the development of an axiomatic approach to deliberative rationality based on the argumentative theory. In Sect. 19.4, we offer a conclusion with suggestions for future work.

19.2 Argumentative Agents: A Semantics for Reasoning Based on Argumentation

The argumentative theory of reason is formulated on the basis of experimental evidence which appears to suggest that human reasoning evolved to facilitate efficient argumentation. From the point of view of an individual, it seems that the most significant purpose of reason is not in helping him to arrive at logically correct forms of inference, but to maximize his chance of winning arguments. In terms of rational choice terminology, one might express this by saying that agents' utility functions are heavily influenced by their desire to do well when they argue with others, often to the detriment of sound classical reasoning.

This insight is interesting in itself, but it is also connected to a second insight stemming from the argumentative theory, one which addresses the very foundation for our understanding of rationality. In particular, the argumentative theory suggests looking for rationality principles that do not target individual reasoners at all, but rather the deliberative processes that they partake in.[4] It is striking how deliberation

[4]This stands in contrast to most formal work on rationality and interaction, which tends to be based on the assumption that agents are individually rational in some appropriate sense, for instance because they seek to maximize given utility functions. Here, we will argue that in order to provide adequate formal foundations for rational interaction we must depart from the approach of trying to reduce it to individual attempts at utility-maximizing. Instead, we propose an approach based on an enactive view of reasoning as an argumentative process, where we model argumentative deliberation as it unfolds along a temporal structure in response to agents' deliberative actions. We mention that models of agents acting in an environment have received much attention from the formal logic community recently, see e.g., Broersen (2011), Alur et al. (2002), Ågotnes and

can often lead to classically sound outcomes even if each individual reasoner is argumentative or even unreasonable, and we follow Mercier and Sperber (2011) in thinking that this mechanism is crucial. It can serve to explain why argumentative reasoning has proved so successful for the human race, even if it regularly leads to unsound, or even absurd, results, when people reason in isolation.

To develop a formal account of argumentative deliberation we will start from a formal representation of the reasoners themselves. In this regard it is important to note that the argumentative theory involves a notion of argumentation which is conceptually distinct from the one usually studied in modern argumentation theory, developed after the influential work of Toulmin (2003) (first edition from 1958). In this research tradition, the focus tends to be directed towards recognizing and categorizing fallacies, as well as the design of argumentation schemes that are meant to facilitate sound reasoning, particularly regarding what arguments we should accept in a given scenario.[5]

It is clear that Mercier and Sperber (2011) asks us to adopt a more descriptive approach, which will allow us to embrace the idea that reasoning evolved as a mechanism to facilitate *efficient* argumentation. This, in turn, entails acceptance of the fact that argumentative reasoners might not conform to any general standards of correctness. We should not model reasoners as what we think they should be, but as what they are: individually unique contributors in a social environment, who often seek to maximize their influence in that environment.

To formally represent argumentative reasoners we will use argumentation frameworks, first introduced in Dung (1995). These are simple mathematical objects, essentially directed graphs, which facilitate the investigation of a whole range of semantics (Baroni and Giacomin 2007).[6] We think they are well suited as a technical

van Ditmarsch (2011), and Hoek et al. (2007). By drawing on argumentation theory, we believe it is possible to develop this work in a direction that will make it even more relevant to the task of modeling situated social interaction.

[5]Some argumentation scholars have criticized this approach, by pointing out that negotiations over meaning is a key aspect of argumentation that most theories developed in the normative tradition seem incapable of accounting for in an appropriate manner, see e.g. Kock (2007) and Wohlrapp (1998). To some, the challenge lies with the inherent *subjectivity* of argumentation, which calls for a shift of focus towards *rhetoric* (Kock 2009). However, others have stressed how argumentation gives rise to important mechanisms whereby people may change their positions and collectively develop novel (and normatively sound) views on the matter under consideration (Wohlrapp 1998). In our opinion, this is the crucial insight, which is also important in relation to the results reported in Mercier and Sperber (2011). In particular, we think it serves as a link between the descriptive and normative content of this theory, as well as with previous work on argumentation.

[6]The theory of argumentation frameworks has been influential in the context of artificial intelligence (Rahwan and Simari 2009). It is capable of capturing many different semantic notions, including semantics for multi-valued and non-monotonic logics, logic programs and games (Dung 1995; Dyrkolbotn and Walicki 2014). The work of Brewka et al. (2011), on the other hand, shows how argumentation frameworks can be used to provide a faithful (and computationally efficient) representation also of semantics that are formulated with respect to the more fine-grained formalism of abstract dialectical frameworks (Brewka and Gordon 2010). It is also important to note that much recent work focuses on providing logical foundations for the theory, work we

starting point towards logics for argumentative deliberation, but we propose to make use of them in a novel way, to represent the agents' subjective interpretations of semantic meaning.[7]

The theory of argumentation frameworks provides us with a flexible formalism for modeling many different ways in which individual agents may choose to reason, using one among the many argumentation semantics that have been developed. Hence, we make no commitment to a given set of reasoning rules; argumentative agents are characterized by the fact that they maintain an argumentative interpretation of the world, not by the fact that they reason about it in a certain way.

19.2.1 Argumentation Frameworks, Agents and Semantic Views

In this section we formally define a logic for argumentative agents, starting with an introduction to argumentation frameworks. Given a set of semantic atoms Π, which we will tend to think of as names of arguments, an argumentation framework (AF) over Π is a relation $E \subseteq \Pi \times \Pi$. Intuitively, an element $(x, y) \in E$ encodes the fact that x attacks y. We can depict E as a directed graph, giving a nice visualization of how the atoms in Π are related as arguments, see Fig. 19.1 for an example. We introduce the notation $E^+(x) = \{y \in \Pi \mid (x, y) \in E\}, E^-(x) = \{y \in \Pi \mid (y, x) \in E\}$, extended to sets $A \subseteq \Pi$ such that, e.g., $E^+(A) = \{y \in \Pi \mid \exists x \in A : y \in E^+(x)\}$. We use $\Pi(E) = \{x \in \Pi \mid \forall y \in \Pi : \{(x, y), (y, x)\} \cap E = \emptyset\}$, denoting atoms from Π that do not appear in any attack from E.

Given an AF E, the purpose of an argumentation semantics is to identify a collection of acceptable sets of arguments, typically called *extensions*. For instance, if $E = \{(p, q), (r, p)\}$, then the semantics might prescribe $\{r, q\}$ as a set that can be accepted, since r defends q against the argument made by p and r is not in turn attacked.

Fig. 19.1 An AF E such that $\Pi(E) = \{p, q, q', p'\}$

can draw on when we develop multi-agent extensions (Grossi 2010; Arieli and Caminada 2013; Caminada and Gabbay 2009).

[7] In terms of each individual agent, using terminology from cognitive science, this means that we employ argumentation frameworks to describe (parts of) the informational level of cognitive processing, see Stenning and van Lambalgen (2008, p. 348) for an informal definition of this term. Previous work has demonstrated that logical tools can have a particularly crucial role to play in facilitating exploration at this level, also serving to shed new light on established truths arrived at through empirical work, see Stenning and van Lambalgen (2008, pp. 348–360), and Stenning and van Lambalgen (2005) for a concrete example. We note that our use of argumentation frameworks to model information-processing and representation makes good sense with respect to the argumentative theory; since agents reason to win arguments, it is natural to assume that they tend to represent semantic information in argumentative terms.

Given an AF E and an extension $A \subseteq \Pi$, the associated three-valued assignment is $\mathsf{c}_A : \Pi \to \{1, 0, \frac{1}{2}\}$ given by

$$\mathsf{c}_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \in E^+(A) \\ \frac{1}{2} \text{ otherwise} \end{cases}$$

This representation corresponds to an intuitive reading where atoms in A are regarded as true/successful, atoms attacked by one of these are regarded as false/failed, and all others are undecided. The three-valued representation facilitates elegant definitions of various argumentation semantics, specifying sets of assignments rather than extensions. In this way, it also becomes natural to reason about AFs using three-valued logic, an idea that has been explored in some recent work Dyrkolbotn and Walicki (2014), Arieli and Caminada (2013), and Dyrkolbotn (2013). This will be exploited in the coming sections, as we will rely on three-valued Łukasiewicz logic when we evaluate complex formulas over given AFs.

All semantics for argumentation of which we are aware are based on the notion of *conflict-freeness*, encoding that it is not allowed to assign 1 to two atoms in conflict. Formally, we say that c is conflict-free for E if the following holds, for all $x \in \Pi$.

$$\mathsf{c}(x) = 0 \iff \exists y \in E^-(x) : \mathsf{c}(y) = 1 \tag{19.1}$$

We use $\mathbf{C}(E)$ to denote the set of conflict-free assignments for E, and we define $\mathsf{c}^1 = \{x \in \Pi \mid \mathsf{c}(x) = 1\}$, $\mathsf{c}^0 = \{x \in \Pi \mid \mathsf{c}(x) = 0\}$ and $\mathsf{c}^{\frac{1}{2}} = \{x \in \Pi \mid \mathsf{c}(x) = \frac{1}{2}\}$. In Fig. 19.2 we provide definitions of the most commonly known semantics based on argumentation frameworks, all based on restricting $\mathbf{C}(E)$.

The logics we introduce are parameterized by the choice of an argumentation semantics, and which we use is not crucial for our analysis in this paper. We note, however, that the *admissible* semantics encode what seems to be minimal criteria for intuitive acceptability of arguments. In addition to requiring that c^1 must be free from internal conflict, it also demands that it must be able to defend itself against all attacks. The semantics from Fig. 19.2 are all based on this idea, but the others involve various additional requirements, all meant to reduce the number of undecided arguments. In the following we use ε refer to some generic semantics from this list.

Fig. 19.2 Various semantics, defined for any $E \subseteq \Pi \times \Pi$

Admissible:	$a(E) = \{\mathsf{c} \in \mathbf{C}(E) \mid E^-(\mathsf{c}^1) \subseteq \mathsf{c}^0\}$
Complete:	$c(E) = \{\mathsf{c} \in \mathbf{C}(E) \mid \mathsf{c}^1 = \{x \in \Pi \mid E^-(x) \subseteq \mathsf{c}^0\}\}$
Preferred:	$p(E) = \{\mathsf{c}_1 \in a(E) \mid \forall \mathsf{c}_2 \in a(E) : \mathsf{c}_1^1 \not\subset \mathsf{c}_2^1\}$
Semi-stable:	$ss(E) = \{\mathsf{c}_1 \in a(E) \mid \forall \mathsf{c}_2 \in a(E) : \mathsf{c}_1^{\frac{1}{2}} \not\supset \mathsf{c}_2^{\frac{1}{2}}\}$
Stable:	$s(E) = \{\mathsf{c} \in a(E) \mid \mathsf{c}^{\frac{1}{2}} = \emptyset\}$

Our first step towards a logic for reasoning about AFs is to define the simple propositional language $\mathscr{L}$, using the grammar in Eq. 19.2.

$$\alpha := p \mid \neg\alpha \mid \alpha \to \alpha \tag{19.2}$$

where $p \in \Pi$. We also use $\alpha \vee \beta := (\alpha \to \beta) \to \beta$ and $\alpha \wedge \beta := \neg(\neg\alpha \vee \neg\beta)$. The three-valued assignments give rise to evaluations of arbitrary formulas from $\mathscr{L}$, defined as in Łukasiewicz logic below.

$$\begin{aligned} &\overline{\mathsf{c}}(p) = \mathsf{c}(p) \text{ for } p \in \Pi \\ &\overline{\mathsf{c}}(\neg\alpha) = 1 - \overline{\mathsf{c}}(\alpha) \\ &\overline{\mathsf{c}}(\alpha \to \beta) = min\{1, 1 - (\overline{\mathsf{c}}(\alpha) - \overline{\mathsf{c}}(\beta))\} \end{aligned} \tag{19.3}$$

This evaluation behaves like classical logic on the semantic values $\{0, 1\}$. For an intuition of how the third value is dealt with, notice that the definition ensures that $\mathsf{c}(p) = \frac{1}{2}$ if, and only if, we have $\overline{\mathsf{c}}(p \leftrightarrow \neg p) = 1$. So we can express that a given atom is undecided, and this is what makes it so natural to use Łukasiewicz logic, as opposed to some other three-valued formalism.

We are now prepared to offer our definition of an argumentative agent. To this end, let $\mathscr{A}$ be a set of agent names. Then a *view* for agent $a \in \mathscr{A}$ is an AF $V_a \subseteq \Pi \times \Pi$. It encodes his interpretation of the meaning of the arguments under consideration. We also define an *argumentative state* as a tuple $(V_a)_{a\in\mathscr{A}}$, associating a view with each agent. In this paper, we will assume for simplicity that the argumentative state remains the same throughout the course of deliberation, so that the views of the agents are not themselves subject to revision as the debate unfolds. This assumption should be relaxed in future work, by nested application of the ideas we develop in the next sections.

Given an agent $a \in \mathscr{A}$ with a view V_a, we can use a modality $\blacklozenge_a$ to perform meta-reasoning about the acceptance status of arguments on AFs, under some arbitrary semantics ε. This idea lead us to the following multi-agent language $\mathscr{L}^{\blacklozenge}$.

$$\phi := \blacklozenge_a\alpha \mid \neg\phi \mid \phi \wedge \phi$$

where $\alpha \in \mathscr{L}$ and $a \in \mathscr{A}$.

Given an argumentative state $\mathscr{B} = (V_a)_{a\in\mathscr{A}}$, we can now define truth for formulas from $\mathscr{L}^{\blacklozenge}$ inductively as follows, for all formulas ϕ.

$$\begin{aligned} &\mathscr{B} \models_\varepsilon \blacklozenge_a\alpha \text{ if there is } \mathsf{c} \in \varepsilon(V_a) \text{ s.t. } \overline{\mathsf{c}}(\alpha) = 1 \\ &\mathscr{B} \models_\varepsilon \neg\phi \text{ if not } \mathscr{B} \models_\varepsilon \phi \\ &\mathscr{B} \models_\varepsilon \phi \wedge \psi \text{ if } \mathscr{B} \models_\varepsilon \phi \text{ and } \mathscr{B} \models_\varepsilon \psi \end{aligned} \tag{19.4}$$

Assume we have a single agent a and that his view V_a is given by the AF in Fig. 19.1. Then it is easily verified that the following claims holds for all semantics ε.

- $V_a \models_\varepsilon \blacklozenge_a q$
- $V_a \models_\varepsilon \neg\blacklozenge_a p \wedge \blacklozenge_a \neg p$
- $V_a \models_\varepsilon \blacklozenge_a(\neg q \wedge q')$
- $V_a \models_\varepsilon \blacksquare_a(\neg q \rightarrow (p \leftrightarrow \neg p))$

We can now begin to explore formally what we mean by classically sound reasoning. Since evaluation of complex formulas agrees with classical logic on the boolean values, it is tempting to say that an agent reasons classically if the argumentation semantics he or she uses always returns two-valued assignments. However, the core requirements underlying the semantics from Fig. 19.2 also seem to capture some crucial aspects of what we mean by classical soundness. Most importantly, they all disallow assignments where two arguments in conflict are assigned 1, corresponding intuitively to the law of non-contradiction. Hence, it appears that the stable semantics, which always require assignments that are both admissible and two-valued, is the best candidate we have for a formalization of classical reasoning about AFs.[8]

Interestingly, if the stable semantics captures classical reasoning, we can formally conclude that there are possible interpretations of semantic reality that make such reasoning impossible. Consider, for instance, what happens if some agent regards an argument to be attacking itself. It is easy to see that such an argument does not admit any stable assignment and that it must by necessity obtain the value $\frac{1}{2}$ under all other semantics from Fig. 19.2. In terms of logic, the self-attacking argument satisfies the formula $\blacksquare_a(x \leftrightarrow \neg x)$, expressing that x is necessarily equivalent to its own negation.

Having defined classical reasoning by the stable semantics, we can also attempt to characterize the necessary failure of such reasoning in terms of graph structures for which no stable assignments exists. In fact, the combinatorial problem of when an AF admits a stable set has been analyzed in graph theory since the 1950s, by researchers using a different terminology, in the field of *kernel theory*. The core result from this field, due to Richardson (1953), immediately implies that any finite AF which has no directed odd cycle of attack admits a non-empty set of stable assignments.[9]

[8]There is strong formal evidence supporting the claim that classical reasoning about AFs is captured by the stable semantics, in particular the result that AFs under the stable semantics provide a normal form for theories of propositional logic (Bezem et al. 2012).

[9]This result was also rediscovered in Dung (1995), but kernel theory offers many additional results and techniques, see for instance Galeana-Sánchez and Neumann-Lara (1984). These results can be understood as providing conditions which ensure the possibility of imposing classical standards of reasoning on agents' interpretations of the world, establishing an interesting link between the formalism in this paper and an established subfield of graph theory.

19.2.2 Extended Example: Rain in Bergen

We now analyze a simple example to motivate the need for introducing subjective views, suggesting also some shortcomings of a traditional approach to argumentation in the context of multi-agent deliberation. For a concrete scenario we consider two agents a, b who argue about whether it will rain in Bergen today. Formally, r represents the claim that it will rain and $\overline{r}$ represents the claim that it will not. We first assume that neither of the agents argue any further in favor of their positions. Then their basic claims are the only semantic entities present, and the AF shown on the left below depicts their relationship. So far, the model appears to be an uncontroversial objective representation of the state of affairs. There is not yet any discernible need for introducing subjective views.

On the right above, we show a simple agent-indexed AF which illustrates a naive attempt at introducing agency to the initial AF. We label the two attacks, from r to $\overline{r}$ and from $\overline{r}$ to r, by both a and b to encode that they are *common* to the agents. That is, both agents acknowledge that these attacks are present – the agents agree that they disagree.

The rational outcome, what claim we *should* accept, remains unclear. Both r and $\overline{r}$ seem acceptable, since no further arguments have been made. Indeed, our formal logic agrees – either r or $\overline{r}$ (but not both) can be taken as true, for all semantics from Fig. 19.2.

Assume that a and b begin to argue in favor of their claims. For simplicity we consider only two steps of debate: first a introduces the argument that the weather report says that it will not rain, and then b counters this by announcing that she has seen a puddle on the pavement, suggesting that it will be a rainy day. Let us call the arguments provided by the weather report and the puddle w and p respectively. Then, noting that w is an argument used by a against r and that p is an argument used by b as a retort against w and also, let us assume, directly against $\overline{r}$, a naive representation in the traditional spirit would be to view the debate as progressing from (1) to (3), as depicted in Fig. 19.3.

This is an uncontroversial depiction of the actual utterance made, but it results in a dubious model. In particular, if we look at the AF (3) depicted in Fig. 19.3 above, we can easily conclude that p, r must be assigned the value true, so that $\overline{r}, w$ are both defeated and become false. Hence, it follows that agent b won the debate. But is this really the rational way of looking at things? Intuitively speaking, it seems problematic. Why would a puddle be conclusive evidence, and stronger evidence than a weather report? For an objective audience, a weather report might carry some weight, and so might a puddle, but neither seem particularly conclusive.

We believe the problem is that the representation we gave in Fig. 19.3 is too simplistic, since it fails to include information about the agents' view of each other's

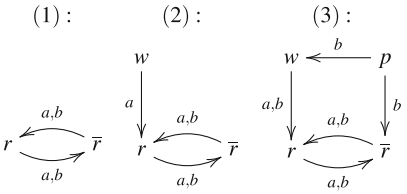


Fig. 19.3 An operational account of the rain debate

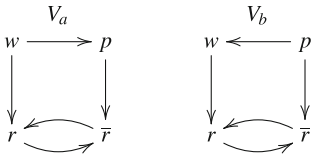


Fig. 19.4 Two views on rain in Bergen

utterances.[10] In particular, what is missing is some account of how agent a views puddles, and what agent b thinks of weather reports. Let us assume that their views are in fact those depicted in Fig. 19.4. These views are clearly consistent with the actual exchange that took place. We also notice that both agents acknowledge that their respective arguments for and against rain are correct. However, they also both think that their own argument is stronger, in that it attacks also the other agent's argument, but not vice versa. This might be the case, for instance, because the weather report gives a reason to doubt that b is telling the truth about the puddle, while seeing the puddle gives b reason to doubt the relevance of the weather report. Importantly, it might not be rational or reasonable for a and b to disagree about how their arguments are related, but that they would do so is nevertheless consistent with the fact that reasoners often tend to display *confirmation bias*, putting more weight on evidence in support of previous beliefs.

Importantly, we can now reject the model presented in Fig. 19.3. The final AF (3), in particular, encodes an interpretation that includes the attack (p, w). But agent a disagrees with agent b about the presence of this attack, so it is unwarranted to take it for granted that (p, q) will come to influence the outcome of deliberation. Indeed, in order to even begin talking about the outcome we need first some aggregated view on the semantic meaning of the arguments involved, based on the agents' views

[10] In a perfect world, this might not matter, since all debate might eventually be settled conclusively by brute empirical fact, such as observing actual rain as opposed to consulting weather reports and puddles. However, in such a world, deliberation would not be very interesting and luckily, deliberation is hardly ever conclusive in the real one. Rather, a debate involves crucially a search for common ground, and common ground depends crucially on how agents perceive the statements made by others, as they reflect on the totality of the debate. This is why we need to be explicit about subjective views and ask for a representation of how each individual agent interprets the semantic meaning of all those claims that are relevant to the scenario at hand.

rather than actual utterances. It seems to us that the function of deliberation is to generate such views, and in the next section we develop a logical framework based on this idea.

19.3 Argumentative Deliberation: A Formalization Using Modal Logic

Given a basis which encodes agents' views of the arguments, we are interested in the possible ways agents may deliberate, and how deliberation can create new interpretations. In short, we want to study the *effect* of deliberation on semantic meaning.[11]

Towards formalization, we first define the set of all possible interpretations that may result. These will be the states of our models, and we will represent them using AFs. At this stage there is only one requirement that we will impose on such states, namely that they are all based on the views of the agents. This is encoded in the following definition.

Definition 19.1. Given an argumentative state $\mathscr{B}$, we say that $q \subseteq \Pi \times \Pi$ is a *deliberative state* for $\mathscr{B}$ if

$$\bigcap_{a \in \mathscr{A}} V_a \subseteq q \subseteq \bigcup_{a \in \mathscr{A}} V_a \tag{19.5}$$

We collect all deliberative states for $\mathscr{B}$ in the set $\mathsf{D}(\mathscr{B})$.

The requirement that states must be based on the agents' views means that we do not allow deliberation to result in states that deviate from interpretations that are held unanimously by the agents. If everyone agrees on the meaning of an argument, the argument has this meaning no matter how deliberation proceeds.[12] Having defined the set of states, we can now define the notion of a *deliberative model*.

[11] There has recently been work done on *merging* of multiple argumentation frameworks, where different agents typically endorse different frameworks, see Coste-Marquis et al. (2007), Gabbay and Rodrigues (2014), and Dunne et al. (2012). This is related to the work we present in this paper. However, merging has so far been studied as an aggregation problem, the goal being to describe principles of aggregation that are "fair" and/or "rational" in lieu of social choice theory (for a list of possible principles, formulated for argumentation scenarios, see Dunne et al. 2012). In our opinion, normative approaches of this kind fail to do justice to the subtle interaction games that take place when agents negotiate about meaning. Therefore, we suggest a different approach, based on branching-time descriptive models of deliberative processes (which may, but need not, result in a merging of different viewpoints), drawing on insights from work done on temporal and strategic logics.

[12] Interestingly, this does not mean that a possible unanimity regarding the semantic status of an argument is necessarily reflected in the view aggregated by deliberation, not even when all agents reason according to the same semantics. If the agents differ in their account of *why* an argument

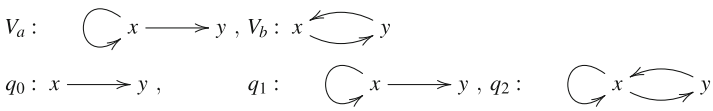


Fig. 19.5 A deliberative model (Q, R) over $\mathscr{B} = (V_a, V_b)$ with $R = \{(q_0, q_1), (q_1, q_2)\}$

Definition 19.2. Given a deliberative state $\mathscr{B}$, a deliberative model for $\mathscr{B}$ is a tuple (Q, R) such that

- $Q \subseteq \mathsf{D}(\mathscr{B})$ is a set of deliberative states for $\mathscr{B}$.
- R is a relation $R \subseteq Q \times Q$.

The idea is that the relation R encodes a process of deliberation based on the views in $\mathscr{B}$. If $(q_1, q_2) \in Q$ the intuition is that there is some event that can take place in the deliberative state q_1 so that the aggregated view is updated, taking us to the deliberative state q_2. We abstract away from the deliberative events that can induce such a link, but this could be some agent presenting his point of view, or it could be some joint effort, say a vote, to reach a decision about some argument. In this paper, we will keep things simple and leave the exact content of events unspecified.

As an example of a deliberative model, consider the framework in Fig. 19.5. Here, the argumentative state is problematic from the point of view of classical logic. In particular, we have $\mathscr{B} \models_\varepsilon \neg\blacklozenge_a x \wedge \neg\blacklozenge_a \neg x$ under all ε from Fig. 19.2, arising from the fact that in agent a's view, the argument x attacks itself and is not defeated. Hence, it cannot be regarded as either true or false without leading to contradiction, and agent a is prevented from reaching any classically sound conclusions about the status of either argument (since he also perceives x to attack y). The agent b, on the other hand, has the view that x and y are in opposition to each other; if one of them is accepted the other must be rejected and vice versa. But she has no information which suggests choosing one over the other. In particular, we have $\mathscr{B} \models_\varepsilon \blacklozenge_b y \wedge \blacklozenge_b \neg y$. Hence, from agent b's point of view, the semantic status of x and y remains unclear. Through deliberation, however, it is possible to arrive at a definite outcome which also resolves the inconsistency that a believes to be present at x.

One such scenario is depicted in Fig. 19.5, where deliberation starts with the framework $\{(x, y)\}$, encoding what the agents already agree on. In this framework, x is the successful argument and y is defeated. Then agent a first puts forth his point of view, resulting in q_1, an anomalous framework where no argument can be either accepted or defeated. But then agent b offers her perspective, resulting in the deliberative state $q_2 = V_a \cup V_b$. Here there is no problem, but now y must be accepted, under all semantics from Fig. 19.2.

This is an example of a scenario where everything runs smoothly and there is no controversy. In particular, both agents uncritically accept adding each others' points

should be accepted, deliberation might lead to its rejection. We will formalize a scenario like this in Sect. 19.3.1, Example 19.3.

of view to the deliberative state, resulting in the union of their views emerging as the final outcome of deliberation. Things might not be so simple, however, and it is the more complicated scenarios that can benefit the most from logical modeling. It could be, for instance, that agent a has reservations about agent b's interpretation of y as an argument that also attacks x. If we are unsure about agent a's stance in this regard, or, more generally, unsure about whether deliberation based on the views of agents a and b will eventually return a state where the (y, x)-attack is included, we can model this by introducing branching in the deliberative model. In particular, we could introduce a reflexive loop at q_1, to indicate the possibility that b's perspective might come to be rejected. Then we have a branching deliberative model, and while it is still *possible* to reach an outcome where x is rejected, deliberation can now also fail to achieve this, if agent b's perspective is not taken duly into account.

To talk about deliberative models in a way that allows us to distinguish and identify situations such as these, we can use existing modal languages of varying expressive power. We will focus on making conceptual points, so we consider the following simple language $\mathscr{L}_1$, which adds to $\mathscr{L}^{\blacklozenge}$ a modality for talking about the current deliberative state and the one-step possibilities in deliberative models.

$$\phi := \blacklozenge\alpha \mid \blacklozenge_a\alpha \mid \neg\phi \mid \phi \wedge \phi \mid \Diamond\phi$$

where $\alpha \in \mathscr{L}$. We use $\blacklozenge$ to talk about deliberative states, while the agent-indexed modalities still apply to the agents' views in the argumentative state. The definition of satisfaction for $\mathscr{L}_1$ on deliberative models is then defined analogously to classical modal logic.

Definition 19.3. Given an argumentation semantics ε, an argumentative state $\mathscr{B}$ and a corresponding deliberative model (Q, R), the truth of $\phi \in \mathscr{L}_1$ on (Q, R) at $q \in Q$ is defined inductively as follows (we omit the boolean cases).

- $\mathscr{B}, (Q, R), q \models_\varepsilon \blacklozenge\alpha$ if $\exists \mathsf{c} \in \varepsilon(q) : \overline{\mathsf{c}}(\alpha) = 1$
- $\mathscr{B}, (Q, R), q \models_\varepsilon \blacklozenge_a\alpha$ if $\exists \mathsf{c} \in \varepsilon(V_a) : \overline{\mathsf{c}}(\alpha) = 1$
- $\mathscr{B}, (Q, R), q \models_\varepsilon \Diamond\phi$ if there is $q' \in Q$ s.t. $(q, q') \in R$ and $\mathscr{B}, (Q, R), q' \models_\varepsilon \phi$

We follow the usual convention of dropping an argument on the left-hand side to signify universal quantification, such that $\models_\varepsilon \phi$ means that for all $\mathscr{B}$, all models (Q, R) for $\mathscr{B}$, and all states $q \in Q$, we have $\mathscr{B}, (Q, R), q \models_\varepsilon \phi$

We define $\Box\phi := \neg\Diamond\neg\phi$ as usual. Consider the model from Fig. 19.5 as an example. It is easy to verify that $\mathscr{B}, (Q, R), q_0 \models_\varepsilon \Box\Box\blacksquare\neg x$, expressing how two steps of deliberation will necessarily suffice to resolve a's semantic problems with x in this scenario, leading us to conclude $\neg x$ at the social level. However, if we add a reflexive edge (q_1, q_1) to this model, to encode uncertainty about whether agent b's view will survive deliberation, we obtain only the weaker $\mathscr{B}, (Q, R), q_0 \models_\varepsilon \Diamond\Diamond\blacksquare\neg x$. It is still *possible* that the problems at x are resolved, but this is no longer necessarily so.

This example shows that we can now formally describe situations where deliberation serves to turn individual views that are problematic into deliberative states that are classically consistent. In the next section, we will see some more examples of how our formalism can be used to capture such mechanisms.

19.3.1 Using Deliberative Logic to Model Argumentative Deliberation

In Mercier and Sperber (2011), one of the primary claims concerns confirmation bias, the mechanism by which reasoners disproportionately tend to favor reasons that support previous beliefs rather than challenge them. According to the authors, this bias is not necessarily an example of flawed reasoning since it has an argumentative function that can serve to enhance the positive effects of deliberation. This claim is supported by empirical evidence, and in this section we show how scenarios where cognitive bias plays a constructive role can be represented by deliberative models and reasoned with using modal logic. Following this, we go on to consider some more examples which we believe illustrate that as an approach to modeling, the formal framework suggested in this paper appears to be both flexible and expressive.

Example 19.1 (Rain in Bergen revisited). We return to the Bergen rain example, considered in depth in Sect. 19.2.2. There we argued that instead of directly modeling the actual exchange of arguments, the deliberative events that took place, we should start from a representation of the agents' view of the arguments. These were given in Fig. 19.4, and we argued that these views were consistent with the deliberative event under consideration. Now we can try again to model the deliberation that took place, more abstractly by seeing it as a traversal on a deliberative model. To do this, we will consider the possible effect that each utterance could have on the deliberative state, given a starting point where r and $\overline{r}$ are in recognized mutual opposition to each other. We recall that the first event was that agent a pointed out the weather report, and that the second event was agent b pointing to the puddle. This leads us then naturally to consider the deliberative model (Q, R), depicted in Fig. 19.6.

For an intuitive presentation of the scenario, we assume that (Q, R) describes a situation where deliberation is based on searching for agreement about what attacks *not* to include. That is, unless the agents agree on something else we end in the deliberative state which is formed by taking the union of the agents' views. Hence, if disagreement runs deep, the Bergen rain debate will end in state q_5. Here it is not hard to see that the question of whether or not it will rain remains unsettled – both $\{w, \overline{r}\}$ and $\{r, p\}$ are admissible in the resulting AF. In this case, then, debating only served to establish the social fact that the question of whether it will rain in Bergen is still open in the social group $\{a, b\}$; both can make up his own mind and disagreement may persist. However, if the agents are willing to

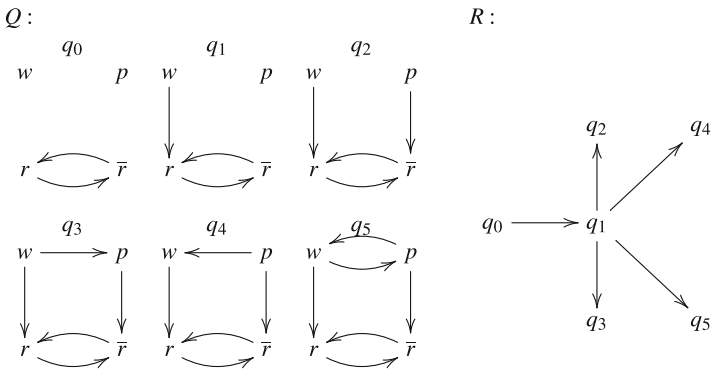


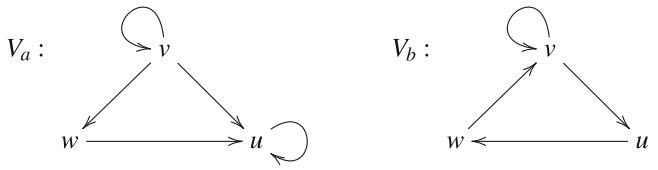
Fig. 19.6 The deliberative model (Q, R) based on the argumentative state $\mathscr{B}$ defined in Fig. 19.4

consider a consensus, then they can settle on either r or $\bar{r}$, by moving to state q_3 or q_4 respectively. Moreover, they can also choose to conclude, in agreement, that the available evidence is *insufficient* to draw any conclusion. This is the outcome resulting from the following deliberative state, which emerges from debate if the agents are prudent and reach agreement on including only those attacks that are present in both views. This means that they end in state q_2, and here both w and p are regarded as successful, meaning that *both* r and $\bar{r}$ becomes defeated and impossible to accept.[13]

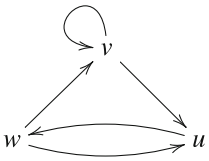
This consensus outcome is interesting since it seems plausible that in an actual debate, this is what one would get if the agents respect each others' input. Hence, it might be the outcome we *should* get. After all, it represent both a clear conclusion, and also a "fair" one in light of the evidence, where neither agent looses on grounds that he or she believes to be unreasonable. More interesting still, it is a consequence of the model that this outcome is only achievable because the agents display confirmation bias with respect to their own arguments; logically, there is little doubt that the two arguments, pulling in opposite directions, attack each other. Yet from the fact that each agent underestimates his opponent's arguments beyond what is rationally warranted, a situation is created whereby deliberation may result in a non-trivial, reasonable interpretation that produces an unambiguous and fair outcome to end the disagreement regarding whether or not it will rain in Bergen today. There is, as usual in Bergen, no way of knowing.

Example 19.2 (Two Wrongs that Make a Right). Consider the argumentative states depicted below.

[13] In this paper we only sketch a framework that permits us to logically examine spaces of possible outcomes, such as those identified by (Q, R). We remark, however, that a natural next step is to try to investigate which one of these would actually result from cooperation, given some assumptions about the faculties of the agents involved, and depending on how arbitration takes place inside coalitions.

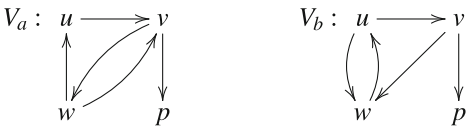


Here the two agents both have an inconsistent view on the semantic elements, as the reader may easily verify. Moreover, the agents agree that v attacks itself. But even so, deliberation can result in consistency being regained. In particular, the AF depicted below is a deliberative state for (V_a, V_b) and it is easily seen to be classically consistent, under the evaluation $\{v \mapsto 0, w \mapsto 1, u \mapsto 0\}$.

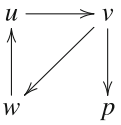


In fact, we can say more about deliberation based on (V_a, V_b). It is true, in particular, that we can evaluate the three arguments classically if, and only if, we end up in a deliberative state where w is understood to attack both v and u. How to ensure that deliberation leads to this result remains unclear, but we can recognize it as a possible way in which, for this argumentative state, two wrongs could in fact make it right.

Example 19.3 (An Agreement that Disagrees with itself). While deliberation can sometimes take us from an inconsistent argumentative state to deliberative states that admit classical evaluation, the direction of deliberation can also go the other way. Consider, for instance, the following views



For $\mathscr{B} = (V_a, V_b)$ we have $\mathscr{B} \models_\varepsilon \blacksquare_a(\neg v \wedge p) \wedge \blacksquare_b(\neg v \wedge p)$. That is, the agents both have an interpretation of the world such that v comes out false and p comes out true under all labellings permitted by any of the semantics from Fig. 19.2. Moreover, both views can be evaluated classically, as the reader may easily verify. But the agents disagree about something else, namely the status of u and w. In particular, we have $\mathscr{B} \models_\varepsilon \blacksquare_a(w \wedge \neg u) \wedge \blacksquare_b(\neg w \wedge u)$. This disagreement could spell disaster for deliberation taking place on the basis of $\mathscr{B}$. It is easy to see, in particular, that the following state q could result through deliberation.



In this case there are *no* classical evaluations available. That is, for any deliberative model containing q, we have $\mathscr{B}, (Q, R), q \models \neg\blacklozenge x \wedge \neg\blacklozenge\neg x$ for all $x \in \{o, v, w, p\}$. Deliberation, in this case, took unproblematic views and produced a problematic outcome. How to avoid this mechanism in general is a difficult problem that deserves further study, but we can already conclude that our model allows us to capture how deliberation can sometimes end up undermining agreements about the status of arguments, and can also lead to the introduction of *new* inconsistencies. Indeed, this too is a mechanism that is intuitively clear. For instance, if one man happens to be right, for all the right reasons, while everyone else disagrees with him, either regarding his conclusions or his reasons, then social deliberation tends to lead to suboptimal outcomes. This is important to remember, particularly since the other, more positive potential in deliberation, is essentially just an expression of the same mechanism. The whole is more than the sum of its parts, and in order to explore and evaluate it we need new tools and principles that sees it as such.

19.3.2 The Search for Formal Characterizations of Rational Argumentative Interaction

Having set up a logical framework for describing argumentative deliberation, we are now also in a position to begin exploring formal representations of principles of rational argumentative interaction. We can think of such principles as restrictions on the class of deliberative models we consider permissible, allowing us to distinguish the "good" scenarios from the "bad" ones in a formally precise way. This can be done semantically, in terms of direct mathematical definition of the appropriate classes of models, or axiomatically, by using formulas and schemata from some formal language to express key properties that we believe can serve to characterize good deliberation.

We have already seen how deliberative logic allows us to model scenarios where the outcome is classically sound even if all individual views are inconsistent. The requirement that deliberation should be organized in such a way that it *always* functions in this way might suggest itself as a good candidate for a normative notion of rationality, defined in terms of social interaction. It is a very strong notion, however, and it is potentially problematic also because it is not in fact wholly social. In particular, a requirement to the effect that the outcome of deliberation should always be classically consistent must by necessity also involve restrictions on what individual views we permit agents to endorse.

This is easy to see intuitively. The case of a system with a single agent who believes something absurd, for instance, or a system with many agents where all of them share an inconsistent interpretation of the world are obvious examples. The fact that deliberation alone cannot ensure a consistent outcome in such cases seems clear, and it is an insight that we can now express formally. First we must define an appropriate formalization of classical soundness at the level of deliberation, and how to do this is not obvious. We will discuss some subtleties regarding this later, but for now let us simply consider the following intuitive axiom schema for classically rational deliberation.

$$\blacklozenge\neg(\phi \leftrightarrow \neg\phi) \tag{19.6}$$

If we require it to be true on all models, in all points, we stipulate that for all claims we can express about the model, it should always be possible to reason about this claim in such a way that it is not considered to be equivalent to its own negation. The restriction is perhaps too weak, but at least it appears like a reasonable minimal requirement. We note, moreover, that the scheme is *not* valid on the class of all deliberative models. Hence, it captures a non-trivial principle, a genuine restriction on deliberation. However, we also notice that for some argumentative states $\mathscr{B}$, there are *no* corresponding deliberative models such that Schema (19.6) holds. Therefore, if we impose it as an axiom of deliberation, we also restrict the class of permissible argumentative states. Hence, we have a formal counterpart to the intuition that constraints on deliberation alone is not enough to ensure classical consistency in all circumstances.

This motivates the definition of a special kind of deliberative rationality principle, which can help us to provide more structure to future inquiries.

- **Liberal principles:** Rationality constraints that do not force us to restrict the set of argumentative states that we consider possible.

We refer to those principles that are not liberal, that is, principles that restrict the permissible argumentative states, as *idealistic principles*. An example of a liberal principle is $\neg\Box(\neg p \wedge p)$, expressing seriality of deliberative models. In the context of deliberation it expresses the principle that deliberation should be open-ended, that there is always a deliberative next step (although at some point it might just be an endless repetition of previously visited states). A more subtle example, involving deliberative interactions, is the principle $\blacklozenge\phi \rightarrow \Box\blacklozenge\phi$, expressing that if something is true in a deliberative state it should also be true in all following states. This is not merely a restriction on the kinds of relations that are permitted to arise in deliberative models, but also a restriction on how deliberation is allowed to unfold from the argumentative state. It is easy to see that it is liberal, however, since a single state without successors will always satisfy it. Notice that if we add a loop to such a state, it witnesses to the liberality of the principle which requires seriality plus commitment to previous outcomes; a debate that never ends and can only increase the set of acceptable truths.

For an example of an idealistic principle, notice that Schema (19.6) is idealistic since it excludes certain argumentative states. In fact, we can provide a simple characterization of those argumentative states that are permitted. Let us say that an argumentative state $\mathscr{B}$ satisfies an axiom schema if there is some deliberative model based on this argumentative state for which the schema is true in all deliberative states. Moreover, let us say that an argumentative state $\mathscr{B} = (V_a)_{a\in\mathscr{A}}$ is finite if $\Pi(V_a)$ is finite for all $a \in \mathscr{A}$. Then we have the following result.

Theorem 19.1. *For all semantics ε from Fig. 19.2, a finite argumentative state $\mathscr{B} = (V_a)_{a\in\mathscr{A}}$ satisfies Schema (19.6) if, and only if, there is some deliberative state q for $\mathscr{B}$ which admits $\mathtt{c} \in \varepsilon(q)$ such that $\forall x \in \Pi : \mathtt{c}(x) \in \{1, 0\}$.*

Proof. $\Leftarrow$) Consider the deliberative model consisting only of q. It follows from Eq. 19.3 that Schema (19.6) is true in q and the claim follows.
$\Rightarrow$) We let $A = \bigcup_{a\in\mathscr{A}} \Pi(V_a)$ and form the conjunction $\phi = \bigwedge_{x\in A}(x \vee \neg x)$. Then for all assignments $\mathtt{c} : \Pi \to \{1, 0, \frac{1}{2}\}$ we have $\overline{\mathtt{c}}(\phi) \in \{1, \frac{1}{2}\}$. Moreover, we get $\overline{\mathtt{c}}(\phi) = \frac{1}{2}$ if, and only if, there is $x \in A$ such that $\mathtt{c}(x) = \frac{1}{2}$. So $\overline{\mathtt{c}}(\neg(\phi \leftrightarrow \neg\phi)) = 1$ if, and only if, there is such $x \in A$ with $\mathtt{c}(x) = \frac{1}{2}$. Let q be a deliberative state for $\mathscr{B}$ such that $\blacklozenge\neg(\phi \leftrightarrow \neg\phi)$ is true in q under ε. This means we can choose $\mathtt{c} \in \varepsilon(q)$ such that $\overline{\mathtt{c}}(\neg(\phi \leftrightarrow \neg\phi)) = 1$, meaning $\mathtt{c}(x) \neq \frac{1}{2}$ for all $x \in A$. Notice that $\Pi(q) \subseteq A$ by Definition 19.1. Hence, $\mathtt{c}(x) \neq \frac{1}{2}$ for all $x \in \Pi(q)$ from which it follows that $\mathtt{c}(x) \in \{1, 0\}$ for all $x \in \Pi$.

This theorem illustrates the kind of result we can obtain when we begin to formalize deliberative principles using logic. It also suggests the subtleties involved and shows that classical concepts can take on new and surprising forms. Notice, for instance, how the case of two wrongs that make a right, considered in Example 19.2, is covered by the result. So according to Schema (19.6), such a collection of views will be permitted, even if none of the views in it are classically consistent.

Instead of forbidding such views, Schema (19.6) requires us to exclude some deliberations that they can give rise to. How to implement such a restriction can be a tricky question, and the language of $\mathscr{L}_1$ might be too weak to deliver principles that allow us to get very far in this regard. Moreover Schema (19.6) might be too strong. For instance, consider a scenario where deliberation proceeds in a step-wise fashion, such that one argument is considered at a time starting from the deliberative state which contains no attacks. This protocol itself seems perfectly reasonable, but it now becomes potentially unreasonable to require classical consistency at every state of deliberation. It might make more sense to stipulate weaker notions, for instance that classical consistency should *eventually* hold, after deliberation has had a chance to work. In general, the ability to express that something holds eventually is an important addition to the expressive power of a modal language. In the context of deliberation it allows us to consider a whole range of interesting notions.

In the following we merely sketch some of these, intended to serve as an illustration of the potential for using stronger modal languages to reason about

deliberative models. In particular, we define the modality $\Diamond^*\phi$, intuitively to be read as saying "after finitely many steps, ϕ becomes true". Formally, we let $\Diamond^n\phi$ denote $\underbrace{\Diamond\Diamond\ldots\Diamond}_{n}\phi$ and define satisfaction for $\Diamond^*\phi$ inductively as follows

$$\mathscr{B}, (Q, R), q \models_\varepsilon \Diamond^*\phi \text{ if there is } n \in \mathbf{N} : \mathscr{B}, (q, R), q \models_\varepsilon \Diamond^n\phi \tag{19.7}$$

We also define $\Box^*\phi := \neg\Diamond^*\neg\phi$. This then expresses "always ϕ". With these constructs in hand we can describe many subtly different properties of deliberation, some of which might be seen as candidates for rationality principles.

Let us assume that ϕ expresses some principle which we take to define "good" states in a normative theory. For instance, ϕ could be an instance of Schema (19.6). Now, even if we believe that ϕ captures some essential normative requirement on the outcome of deliberation, it is not clear that we should require *all* states in a deliberative model to be good states. Indeed, it can often seem more natural to think of deliberation as a process that should ideally take us from bad to good states. Then it is inappropriate, and overly simplistic, to implement a normative ideal by simply forbidding bad states. Instead, we might want to restate our principle ϕ in one of the following ways, as a requirement on what it should be possible to achieve through deliberation starting from the current state.

- $\Box\phi$; all deliberative events take us to a state where ϕ is true.
- $\Diamond\phi$; there is at least one event taking us to a state where ϕ is true.
- $\Diamond^*\phi$; there is a chain of events such that ϕ eventually becomes true.
- $\Diamond^*\Box\phi$; there is a chain of events taking us to a state where every event will make ϕ true.
- $\Diamond^*\Box^*\phi$; there is a chain of events taking us to a state where no further chain of events can make ϕ false.
- $\Box^*\Diamond^*\phi$; for every chain of events, there is a way to continue this chain so that ϕ eventually becomes true.
- $\Box^*\Diamond^*\Box^*\phi$; for every chain of events, there is a continuation so that ϕ eventually becomes true, and remains true forever.

These are examples of notions that can be formulated using temporal languages. In the context of classical modal logic they are quite well understood, but when ϕ also involves occurrences of modalities such as $\blacklozenge$, expressing argumentative properties of states, new issues arise, both technical and philosophical. One of the most interesting questions concerns expressive power and in future work we hope to consider various temporal and fixed-point languages, asking what they allow us to say about deliberative models. Moreover, we think it will be fruitful to combine this technical work with addressing the philosophical challenge of developing a better understanding of the nature of argumentative deliberation. To study normative principles axiomatically is a particularly interesting aspect of this work, and we believe the preliminary investigation carried out here shows its promise.

19.4 Conclusion

We have presented a formal approach to the argumentative theory of reason (Mercier and Sperber 2011), and we have argued that this line of inquiry should be pursued further. The argumentative theory is interesting because it suggests an alternative approach to rationality, one which sees it as irreducibly social and emergent from deliberation. Hence, we think it can contribute important insights relevant to the search for a new foundational theory of reason, a theory that could prove relevant to a range of different fields, including artificial intelligence.

We hope the simple framework developed in this paper can provide a point of departure for the development of more sophisticated technical tools. In the first instance, these should be designed so that they allow clear and informative mathematical modeling of important aspects of argumentative deliberation. However, we also think that formal models can facilitate a highly interesting axiomatic approach to deliberative rationality, allowing logical investigation of foundational issues, assisted by the use of temporal and fixed-point modal languages interpreted on argumentative structures. On the technical side, we think the study of the expressive power of various languages that allow interactions between computational and argumentative modalities is highly interesting and should be considered further.

The study of deliberative models is the study of how representations of meaning and communicative events together give rise to branching structures of social reality, where what is truly sound and rational might be hidden somewhere in a combinatorial object, a recurring pattern, or else not emerge at all except in the limit. But wherever it is hiding, and whatever its nature might be, we believe the pursuit itself has merit and should continue. In the end, perhaps the search for truth is itself a deliberative imperative, the normative power of which is derived not from the fact that it may some day settle, but from the fact that it must always be carried on.

References

Ågotnes, T., & van Ditmarsch, H. P. (2011). What will they say? Public announcement games. *Synthese, 179*(Supplement-1), 57–85.

Alur R., Henzinger T. A., & Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM, 49*(5), 672–713.

Arieli, O., & Caminada, M. W. A. (2013). A QBF-based formalization of abstract argumentation semantics. *Journal of Applied Logic, 11*(2), 229–252.

Baroni, P., & Giacomin, M. (2007). On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence, 171*(1015), 675–700.

Bezem, M., Grabmayer C., & Walicki, M. (2012). Expressive power of digraph solvability. *Annals of Pure and Applied Logic, 162*(3), 200–213.

Blume, L., & Durlauf, S. (2001). The interaction-based approach to socioeconomic behavior. In S. Durlauf & P. Young (Eds.), *Social dynamics*. Washington, DC: Brookings Institutions Press.

Brewka, G., Dunne, P. E., & Woltran, S. (2011). Relating the semantics of abstract dialectical frameworks and standard AFs. In T. Walsh (Ed.), *IJCAI*, Barcelona (pp. 780–785). IJCAI/AAAI.

Brewka, G., & Gordon, T. F. (2010). Carneades and abstract dialectical frameworks: A reconstruction. In *Proceedings of the 2010 Conference on Computational Models of Argument: Proceedings of COMMA 2010* (pp. 3–12). Amsterdam: IOS Press.

Broersen, J. (2011). Making a start with the *stit* logic analysis of intentional action. *Journal of Philosophical Logic, 40*(4), 499–530.

Caminada, M. W. A., & Gabbay D. M. (2009). A logical account of formal argumentation. *Studia Logica, 93*(2–3), 109–145.

Coste-Marquis, S., Devred, C., Konieczny S., Lagasquie-Schiex, M.-C., & Marquis, P. (2007). On the merging of Dung's argumentation systems. *Artificial Intelligence, 171*(10–15), 730–753.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence, 77*, 321–357.

Dunne, P. E., Marquis, P., & Wooldridge, M. (2012). Argument aggregation: Basic axioms and complexity results. In B. Verheij, S. Szeider, & S. Woltran (Eds.), *COMMA* (Vol. 245, pp. 129–140). Vienna: IOS Press.

Dyrkolbotn, S. (2013). The same, similar, or just completely different? Equivalence for argumentation in light of logic. In L. Libkin, U. Kohlenbach, & R. Queiroz (Eds.), *Logic, language information, and computation* (Vol. 8071, pp. 96–110). Berlin/Heidelberg: Springer.

Dyrkolbotn, S., & Walicki, M. (2014). Propositional discourse logic. *Synthese, 191*(5), 863–899.

Gabbay D. M., & Rodrigues, O. (2014). An equational approach to the merging of argumentation networks. *Journal of Logic and Computation, 24*(6), 1253–1277.

Galeana-Sánchez, H., & Neumann-Lara, V. (1984). On kernels and semikernels of digraphs. *Discrete Mathematics, 48*(1), 67–76.

Grossi, D. (2010). On the logic of argumentation theory. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 – Volume 1* (pp. 409–416). Richland: International Foundation for Autonomous Agents and Multiagent Systems.

Hoek, W., Roberts, M., & Wooldridge, M. (2007). Social laws in alternating time: Effectiveness, feasibility, and synthesis. *Synthese, 156*(1), 1–19.

Kock, C. (2007). Norms of legitimate dissensus. *Informal Logic, 27*(2), 179–196.

Kock, C. (2009). Choice is not true or false: The domain of rhetorical argumentation. *Argumentation, 23*(1), 61–80.

List, C., & Dryzek, J. (2003). Social choice theory and deliberative democracy: A reconciliation. *British Journal of Political Science, 33*(1), 1–28.

Mead, G. H. (1967). *Mind, self and society* (3rd ed.). Chicago: University of Chicago Press.

Mercier H., & Sperber D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(02), 57–74.

Ossowski, S. (Ed.). (2013). *Agreement technologies* (Vol. 8). Dordrecht: Springer.

Parikh, R. (2002). Social software. *Synthese, 132*(3), 187–211.

Rahwan, I., & Simari, G. R. (Eds.). (2009). *Argumentation in artificial intelligence*. Dordrecht/New York: Springer.

Richardson, M. (1953). Solutions of irreflexive relations. *The Annals of Mathematics, Second Series, 58*(3), 573–590.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence, 94*(1–2), 57–77.

Stenning, K., & van Lambalgen, M. (2005). Semantic interpretation as computation in nonmonotonic logic: The real meaning of the suppression task. *Cognitive Science, 29*(6), 919–960.

Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge: MIT Press.

Terrell, T. P. (2012). The art of legal reasoning and the angst of judging: Of balls, strikes, and moments of truth. *Northwestern Journal of Law and Social Policy, 8*(1), 35–88.

Toulmin, S. (2003). *The uses of argument* (2nd ed.). Cambridge/New York: Cambridge University Press. (First edition from 1958).

van Benthem, J. (2008). Logic and reasoning: Do the facts matter? *Studia Logica, 88*(1), 67–84.

van Benthem, J. (2011). *Logical dynamics of information and interaction*. Cambridge/New York: Cambridge University Press.

Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic, 38*(6), 649–680.
de Waal, F. B. M., & Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. *Trends in Cognitive Sciences, 14*(5), 201–207.
Wohlrapp, H. (1998). A new light on non-deductive argumentation schemes. *Argumentation, 12*(3), 341–350.

Chapter 20
Explaining Everything

David Davenport

Abstract Oxford physicist David Deutsch recently claimed that AI researchers had made no progress towards creating truly intelligent agents and were unlikely to do so until they began making machines that could produce creative explanations for themselves. Deutsch argued that AI must be possible because of the Universality of Computation, but that progress towards it would require nothing less than a new philosophical direction: a rejection of inductivism in favour of fallibilism. This paper sets out to review and respond to these claims. After first establishing a broad framework and terminology with which to discuss these questions, it examines the inductivist and fallibilist philosophies. It argues that Deutsch is right about fallibilism, not only because of the need for creative explanations but also because it makes it easier for agents to create and maintain models—a crucial ability for any sophisticated agent. However, his claim that AI research has made no progress is debatable, if not mistaken. The paper concludes with suggestions for ways in which agents might come up with truly creative explanations and looks briefly at the meaning of knowledge and truth in a fallibilist world.

Keywords AGI • Creativity • Explanation • Mental models • Fallibilism • Computationalism • Truth • Knowledge • Information

20.1 Introduction

Is it possible to explain everything? Doing so would certainly involve explaining how we human beings not only explain the world around us, but also do all the other complicated things we do. Might it involve something more, something beyond even our abilities? These are questions that David Deutsch, an Oxford physicist renowned for his work on quantum computation, attempts to answer in his recent book (Deutsch 2011), "The Beginning of Infinity". He claims that any agent (natural or artificial) endowed with a certain minimum set of capabilities, can begin this

D. Davenport (✉)
Bilkent University, Ankara 06800, Turkey
e-mail: david@bilkent.edu.tr

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_20

infinite journey towards explaining everything. He goes on to suggest, however, that Artificial Intelligence (AI) research has made no progress towards this goal. What's missing, he argues, is the ability to autonomously generate creative new explanations, and he traces this failure to the inductivist philosophy he sees as inherent in much AI work, promoting fallibilism as the way forward.

This paper examines Deutsch's claims, finding some merit in them, but also suggesting that some AI research is indeed moving in the direction he suggests. I begin by considering his claims about AI's lack of progress and the problems we still face with even basic terminology. I then take an in-depth look at the inductivist philosophy that Deutsch claims is to blame for AI's woes, and examine the fallibilist philosophy he proposes we adopt instead. I will argue that the philosophy we choose directly impacts the architecture and implementation of the agent, and that ultimately the fallibilist approach not only makes more sense, but makes it much easier to construct and maintain the internal (mental) models necessary for true intelligence. These considerations may even have implications for philosophy, requiring us to take a fundamentally different view of what constitutes knowledge and truth.

20.2 Progress in AI?

Clearly, we do not yet have the sort of intelligent machines Science Fiction writers and Hollywood films have long envisaged—and perhaps that's a good thing,—but has there really been no progress towards AI, as Deutsch claims? To some extent the answer depends on who you ask, what their expectations are and what they see as the goal of AI research. For some, especially philosophers and neuroscientists, the purpose of AI research is to understand how the human brain works its magic. While I can't vouch for the philosophers, neuroscientists are certainly beginning to develop a good understanding of the (low-level) biological basis of cognition, though they still have a very long way to go. For others, engineers in particular, the aim is usually just to build better, smarter machines. Whether such devices solve problems in the same way that humans do is usually irrelevant; they may be inspired by solutions seen in the human or animal kingdom, but they are just as likely to adopt a completely different approach if they can find one that produces results more efficiently than evolution has thus far offered. Of course, most such devices are restricted in scope. Missile control systems and anti-lock braking systems are just a few examples of literally hundreds of thousands of these mundane systems that incorporate algorithms and techniques developed by AI researchers, many of which perform far better and more reliably than humans could possibly manage. To date, very few machines have matched, let alone surpassed, human-level performance in what most people would consider intellectual tasks—playing chess, diagnosing human illnesses, playing Jeopardy, etc. Those high-profile systems that have, such as IBM's Deep Blue, which beat World chess champion Gary Kasparov in 1997, and its recent successor Watson, which won the TV game show Jeopardy just last year, are far from general-purpose; being focused on a specific task, most are simply

unable to cope with even minor changes to that task or to the environment (though Watson is more flexible than most).

While "clever", none of these machines could be described as being really "intelligent". But, then, what is intelligence? Surprisingly, we do not even have a good definition of intelligence. Humans are the only widely-accepted example of intelligent beings, and even this status is open to debate. Legg (2008) collected and analysed around 70 definitions from the literature and from experts, synthesising his own definition: "Intelligence measures an agent's ability to achieve goals in a wide range of environments." Notice that this is an essentially behaviourist definition, as it doesn't concern itself with how the agent achieves its goals, only that it does achieve them.

Given that the very notion of intelligence is so vague, is it any wonder that progress towards it has been painfully slow? How do we know we are even on the right path, let alone getting closer? IQ tests, as applied to humans, have a long and chequered history (Kaufman 2000). Legg (2008) discusses intelligence tests for humans, animals and machines—the latter including not only the infamous Turing Test, but also compression tests, linguistic complexity tests, multiple cognitive abilities, competitive games, etc.

Real intelligence requires achieving goals "in a wide range of environments", something most AI systems to date have found difficult to achieve. It was this obvious lack of generality that sparked the Artificial General Intelligence (AGI) program. Ben Goertzel, the researcher who coined the term AGI, responded to Deutsch's claims in "The real reasons we don't have AGI yet" (Goertzel 2012), listing: the weakness of current computer hardware, the relatively minimal funding, and the integration bottleneck—that is, "intelligence depends on the emergence of certain high-level structures and dynamics", and "we have not discovered the one algorithm or approach capable of yielding the emergence of these structures", ... "integrating a number of different AI algorithms and structures is tricky" and "so far the integration has not been done in the correct way". All of which sound more like excuses that merely confirm Deutsch's claim about the lack of progress in AI. And yet Deutsch, like Goertzel, passionately believes that AGI must be possible.

> "Despite this long record of failure, AGI must be possible. And that is because of a deep property of the laws of physics, namely the universality of computation. This entails that everything that the laws of physics require a physical object to do can, in principle, be emulated in arbitrarily fine detail by some program on a general-purpose computer, provided it is given enough time and memory. ... [In the 1980s] I proved it using the quantum theory of computation. (Deutsch 2012)

So is Deutsch correct in his analysis? Others in the field have accused him of dabbling in topics outside his area of expertise and of being unaware of recent developments in the field. Whether Deutsch is aware of current research in AI and cognitive systems, I don't know; nowadays it is difficult for anyone to follow research in such a broad field. Even so, we shouldn't dismiss ideas just because they come from an outsider; without the burden of prior knowledge they may recognise things that those "in the trenches" are just too close to see.

Unfortunately, we continue to encounter the same difficulties when trying to define many other terms commonly used in discussing AI, including "information",

"explanation", "representation", "truth", "knowledge", "computer", and "computation". This, surely, is an indication of a field that lacks solid foundations, its practitioners locked in endless wrangling over terminology. In an attempt to prevent such misunderstandings, at least within this paper, I will begin by outlining how I see some of these notions. I will not attempt to give precise definitions, but rather indicate the general ideas and relationships as I see them.

20.3 Agent Architecture, Models and Terminology

As depicted in Fig. 20.1, agents are a part of the world—that is, we can view them as having a physical existence (a body)[1] in a physical environment (the world).[2] Agents have sensors (inputs), that allow them to detect only certain very limited aspects of the world, and physical actuators (outputs), that enable them to bring about (very limited) changes in the world. In between the sensory input and the actuator output stands the agent's "control" mechanism, which connects (maps) the inputs to outputs.

This mapping may be a very crude and direct mechanism that produces the same response whenever a particular input is sensed—for example, the automatic light

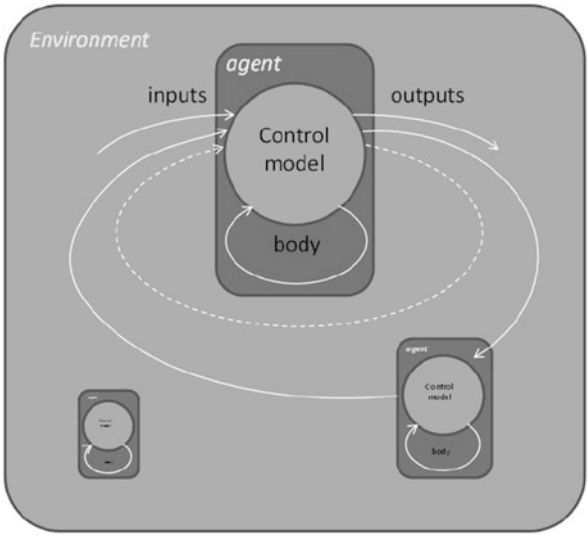


Fig. 20.1 Agents interact with an environment that includes other agents. Each agent has a body and a control mechanism that mediates between its input sensors & output actuators

[1]While agents commonly have well-defined physical boundaries, there is no obvious need for such a restriction, c.f. distributed agents and the extended mind hypothesis.

[2]Nothing in what follows actually depends on there being a physical reality; there is simply no way we can prove that an external world exists, or that it has existed for more than a few seconds prior to the present moment, nor that it is not a figment of the individual's imagination.

inside a refrigerator that switches on when the door is opened, a thermostatically controlled radiator, or Watt's centrifugal engine speed governor. Such devices, while useful, barely qualify as agents. Of more interest, are mechanisms whose mapping involves multiple internal states, and agents that can learn from (i.e. modify the mapping as a result of) their interactions with the environment, such that they may produce different responses in essentially similar circumstances. Agents can vary dramatically in terms of the number of inputs, outputs, and states available for the mapping, as well as the internal organisation/mechanism responsible for managing all this.

The sort of agents we are here primarily concerned with, are ones that can create and maintain internal (mental) "models" of their environment, models they can use to predict the effect of their actions or inaction, so as to enable them to select the output(s) most likely to benefit them. In other words, agents have "goals", that is, they desire to bring about certain states in the world.[3] In the case of biological agents, such goals include satisfying very basic survival needs—for food and water, for flight from predators, for reproduction, etc. If the agent exists in a social group, then more abstract, long-term goals may also be apparent—e.g. those related to job, responsibility, relationships, etc. The agent's task then, is to select appropriate action sequences in order to satisfy its goals as best it can, given its limitations and the vagaries of the world. The degree to which agents succeed in managing their world is a measure of their intelligence (c.f. Legg's definition of intelligence in the previous section). Given the huge variability (seemingly) inherent in the environment, the more sophisticated an agent's ability to "understand" its world (to model and predict it), the more successful it is likely to be.

An agent's model, then, is its "understanding". What is a model? A model is something that allows us to answer questions about another system (the system-being-modelled—the environment, for instance) without actually needing to interact with the other system. Such a substitute is particularly useful when the system itself is too big or too small to manipulate, or when doing so would be too dangerous, or when it doesn't actually exist but we would like to know how it would react were we to build it. Models are useful in everyday situations too: from predicting where a moving object will reappear after being obscured, to calculating how to move one's arm and hand to pick up a cup of coffee. Even very conventional tasks, such as stock-control, can be formulated in terms of models. The model need only correspond to the system-being-modeled in ways relevant to answering the questions that will be asked of it. In most cases, this will require the model have states that can be mapped onto those of the system-being-modeled and that the sequence through which such states evolve from given initial conditions, correspond to those that the system-being-modelled would go through in the same circumstances. The model, then, can be used to simulate the relevant aspects of the system-being-modelled. Of course, such a model may not be entirely accurate and

[3] Agents may not know what these states are, and initially may have them fulfilled either accidentally or purposefully through the actions of other agents, such as parents.

external influences acting on the system-being-modelled may result in it behaving in ways unforeseen by the model (e.g. in the stock-control system, parts may be mislabelled or stolen, leading to inaccuracies in the model). Long term predictions are thus likely to be increasingly error-prone. Notice that the speed of simulation is not especially relevant, what matters—what defines the simulation—is the sequence of states/events.[4] Of course, for an agent interacting with the real world, being able to perform simulations rapidly enough is vital, as it can mean the difference between life and death.

Any physical system with dynamic (causal) characteristics that can be mapped to those of the system-being-modelled, can act as a model. This notion of model is clearly material-independent; that is, models of the same system could be implemented (instantiated) in many different ways and in a variety of different materials, all of which would allow the user to run the same simulations and so answer the same set of questions.[5,6] Models may be constructed either by finding an existing physical system with the appropriate causal structure, by constructing such a system anew, or, more commonly nowadays, by programming a general purpose computer.

The digital stored-program (von Neumann) computer is a universal computing device that can easily be set up ("programmed") so as to model (almost) anything. A program (algorithm), then, is an abstract specification for a causal system that can be used as a model, represented in a language and form that the machine can interpret so as to set up the appropriate causal connections. A program (algorithm) defines a model and so a set of causal systems (computers) that would implement the model. A program transforms a universal machine into a specific single purpose machine. We also commonly say that a program specifies a computation, meaning either a single run of the model with given inputs or the entire set of runs of which it is capable. Whilst the term "computer" has become practically synonymous with the general purpose von Neumann machine, it should now be clear that the term is much more general, applying equally to general and specific purpose machines howsoever constructed. Deutsch sees the "universality of computation" as a "law of physics", which seems a strange categorisation given that computation is specified independently of the implementation (and that, as we will see shortly, the very notion of a "law" is at odds with what Deutsch is proposing). His claim that any physical system "can be emulated in arbitrary fine detail" seems correct, though for the wrong reason, since computation need not be restricted by time and memory as are machines of the von Neumann ilk—another instance of the same physical substance could be used to perform the computation.

Returning to Fig. 20.1, the agent's control system (a computational model) affects the body, which in turn affects the environment; the loops—the feedback

[4]What the states are, how they are mapped, and how they are recognised and interpreted are also important questions.

[5]Of course, the model's states and their interpretation may well change between implementations.

[6]Not every material is necessarily suitable.

that results in subsequent input to the control system—may occur wholly within the body, and/or directly or indirectly via the environment, including other agents. This "picture" shows that the embodied, embedded and situated approaches to cognition merely emphasise different aspects of the same basic arrangement, while all remaining computational in nature (i.e. Computationalism is "Still the only game in town" (Davenport 2012a, b), given the understanding of computation outlined above).

One very important step on the path to explaining everything is the acquisition of language. Agents can use their physical outputs to generate signs/symbols in the form of sounds and/or the placement of materials, in such a way that these come to have meaning for other agents and for the agent itself. The inputs an agent receives can then include such audio and visual linguistic elements, as well as the "raw" input from the environment. We can now define "information" as input that "forms" the model or input that "informs" the model. That is, an agent's input contributes to the model (as experiences) and is used by the model to decide what is "out there" and hence how to react. Such linguistic elements enable the agent to bring to mind things which are not currently in their immediate environment (something especially useful once the agent has developed an inner-voice) effectively freeing them from the here-and-now and allowing the conjunction of arbitrary concepts. The outputs an agent makes now also include linguistic elements, in the form of commands, requests, questions, answers, descriptions, and explanations, that induce other agents to make changes in the environment or ask them for information so as to update the model. Other elements, however, are intended to communicate the contents/structure of the agent's model, so as to update the mental models of other agents.

Note that we often speak of the information being contained in the model and of the model being the explanation for some phenomena. Deutsch points out that prediction is not explanation—you can predict that the woman sawn in half by the magician will be unharmed, but what you really want is an explanation of how the trick works and why your observations fail you. Thus, when Deutsch uses the term "explanation" he is actually referring to the knowledge inherent in the agent's model, rather than the spoken statement of its content. This usage is understandably common in scientific circles, since science aims to provide an objective picture of a phenomena; a scientific explanation refers, in essence, to the "knowledge" that is common to the models in the heads of all scientists. Woodward (2011) provides a good overview of theories of scientific explanation.

Given this general outline of cognitive agents as being ones that employ computational models to maximise their chances of success, the difficult technical question of just how a physical mechanism can create, maintain and use such models, arises. There has been no shortage of suggestions from philosophers and the AI community, as to how to do this, yet to-date there seems little consensus regarding the proper approach. It is here, I believe, that philosophy really matters and that Deutsch's claim that much work in AI has adopted the wrong, "inductivist", philosophy makes sense. Deutsch argues that if progress is to be made, AI must instead adopt "fallibilism" in order to be able to generate novel hypotheses. I agree,

and I also believe that fallibilism will make it easier to create and use models. To understand why, we first need to examine both of these philosophies.

20.4 Inductivism Versus Fallibilism

Induction was (and in many ways still is) seen as fundamental to science and the scientific method. It was part and parcel of the romantic view of science and scientists discovering Knowledge and revealing the Truth about Nature, that has long held sway (at least in Western thought). Induction, in essence, involves taking a set of individual observations and constructing a universal rule (or law) from them.[7] The paradigmatic example involves deriving from the observations that the Sun rose in the East today, and yesterday, and the day before that, and indeed for all of recorded history, the rule that, "the Sun rises in the East every day" (hence justifying the inference that it will rise in the East tomorrow). Similarly, the observation that people get old and die, allows us to induce the law that, "all men are mortal" (so that, given "Socrates is a man", we can deduce that "Socrates is mortal"). Another classic example takes observations of lots of white swans, to derive the rule that, "all swans are white". While the former examples appear reasonable, the latter shows the weakness of this form of reasoning. That it is invalid becomes apparent the moment you encounter your first black swan (and is further eroded when you learn that there are actually swans which are both black and white). The problem, as pointed out long ago by Hume, is that it is simply impossible to examine all relevant instances, as would be necessary to guarantee a law-like relation. The fact that induction is demonstratively invalid thus presents a very serious problem for a scientific method that seemingly relies on examining the world and discovering such universal laws (Vickers 2013).

It was Karl Popper (1959) who pointed out that no matter how many matching instances were uncovered, scientific "laws" were actually just theories that could never be confirmed (proven correct), but that could be falsified. A single failure to match (like the discovery of a black swan), could in principle invalidate a scientific theory. In practice, of course, scientists do not immediately eliminate a theory the moment they come across a single mismatch. Rather, they will first look to explain the erroneous value; were the experimental conditions wrong, were the measurements mistaken, etc. Repeatability of experiments is critical to the validity of science. Even if the failures were real, scientists would tend to hold on to an existing theory, embellishing it with special cases as necessary, until it would eventually "collapse under the weight of its own improbability". And even then, until a new theory could be found—one that covered all the troublesome cases as

[7]Some people include abduction and other uncertain inferences under the general heading of induction. In this paper I use induction in the narrow sense outlined above, rather than adopting this broader sense.

well as the "normal" cases—the existing theory would continue to hold sway. The notion of a critical experiment, one that clearly falsified a theory and so forced practitioners to switch to a new paradigm, was a convenient fiction—a rewriting of history. Rather, as subsequent studies in the Philosophy of Science (by Lakatos, Kuhn and Feyerabend, among others) have demonstrated, scientific progress is much messier and more subject to social forces than many, including Popper, had realised. The idea that multiple theories can usefully co-exist is particularly significant, especially considering that none of them need necessarily be complete or even correct. Indeed, it becomes clear that the very notion of a single completely correct theory may be an unattainable ideal.

None of this, however, detracts from Popper's insight into the nature of scientific theories; the fact that they are fallible, and forever subject to revision. It is this fallibility, coupled with the idea that all scientific theories are actually conjectures that can and must be tested and replaced if found wanting, that Deutsch sees as necessary for progress towards true AI. The question of how and whence conjectures come, is now of crucial importance. Indeed, part of Deutsch's claim, is that creativity and, in particular, the ability to create new explanations (new conjectures), is what is necessary for agents to be able to explain everything. This seems correct. Consider Deutsch's example, the explanation that the points of light we see in the night sky are actually very distant suns. Surely no amount of induction would generate such an idea. It requires an intellectual leap—a flash of inspiration—of the sort that happens so rarely, that the humans who actually perform it (and are able to convince others of it) are often regarded as geniuses.

20.5 Model Building Made Easier?

The task of an individual agent parallels that of the scientist. As we have already seen, agents sense the world (make observations) and use this information both to form models, and as input for existing models which then generate predictions to guide future actions. If we already had a completely correct model, this latter task would be relatively easy. The problem is that we don't, and acquiring it in the first place corresponds to inductivism, which is impossible without infinite time, a luxury that real world agents most definitely do not have.

The alternative, fallibilism, places no such constraints on the correctness of its model, indeed, it isn't even restricted to a single model. Fallibilism allows multiple, possibly incomplete, possibly erroneous models to coexist. What matters—what is absolutely crucial to fallibilism—is that the "incorrect" models can, sooner or later, be removed from consideration leaving the better ones to guide the agent's future actions. This is Darwinian "survival of the fittest" (most useful) at an altogether different level. Two questions now arise: how can an agent generate new hypotheses, and, how can it select relevant models and weed out the "bad" ones?

How can an agent know which hypotheses are the "bad" ones, how can it know it is wrong? In one sense it can't. All it can do is predict what it might expect to happen

in a given situation and then sense what actually happens. If what happens is what it expected, its model seems sound, but if something else—something unexpected—occurs then clearly its model is not completely correct. It can't immediately know what is wrong with its model, but it does now have additional information about the situation, which it can take into account in the future. At higher levels (i.e. levels involving language and abstract thought), it is possible that the agent could run some simulations "in its head", determine whether or not the results were consistent with its expectations, and "label" the hypotheses accordingly so that they are easily included or excluded from future consideration. Only those hypotheses which appear reasonable need be tested.

This, then, provides an insight into how agents can construct hypotheses/models in the first place. Having remembered (stored) as much as they can of what happens, including any actions they make, they match the new (sensory) inputs against the existing ones. The process of doing this is essentially abductive inference, guided by top-down expectations from existing hypotheses[8] (the memories created when the agent first encountered something different). Those hypotheses (memories) that recur and so prove useful, live to fight another day, whilst the others may eventually die off or at least not be included in subsequent computations. This, of course, is not enough for AGI, but it does get us started. It only takes a few small miracles to get from here to the other major component we need, which is language. Once an agent can name (and so group) arbitrary concepts, it can more easily detach itself from the here-and-now and contemplate currently unsensed things. Much of what constitutes (high-level) creativity is simply the result of chance combinations of the present sensory experience with memories (knowledge) of things previously named. It is this simple "conjunction" (joining) of ideas that is the spark. Some prolific inventors claim to adopt this approach, systematically combining concepts at random and investigating the consequences.

Of course there are other, more sophisticated ways in which agents can generate new hypotheses, including, abstraction, analogy and what we might refer to as pseudo-induction. By pseudo-induction I mean extracting "rules" from collections of observations, exactly as with induction, but with the "understanding" that the results are merely hypotheses not universal laws. This then allows us to continue to make use of induction without the epistemic worries, which is exactly what scientists and most of the machine learning community actually does.

Analogy is obviously very closely related to modelling and explanation, and was also proposed in a response to Deutsch by Wiedermann (2013). Analogical reasoning comes in many varieties (Bartha 2013). In essence, it involves selecting corresponding objects/states/processes between two systems (or perhaps more correctly, between models of two systems)—one of which we are usually familiar with—, noting that most of these relations hold true and from that inferring that others, whose status is currently unknown, are also true. Of course, there is no

[8]There is an exact parallel here with the idea from the Philosophy of Science that all experimentation is carried out within the context of a particular theory.

obvious reason that this should be so, hence analogy, like induction, is an invalid form of inference. It is, however, very common. In one sense we use analogies whenever we do something in a slightly different context. For example, each restaurant we visit is different, but we can abstract and draw analogies between them, such that we are usually successful in getting the food we want. More interesting and creative forms of analogy involve mappings between two completely different domains. A classic example of this would be modelling atomic particles as billiard balls. This works fine, until we realise that in some situations particles actually behave more like waves (another analogy) than billiard balls. Unfortunately, physicists were unable to find an analogy that provided the necessary intuitions, leaving them with the so-called, wave-particle duality. A similar situation is now being played out in quantum physics, for which there seems no suitable analogy at all.[9]

Finally, we should not forget that explaining everything should include not only what exists, but also what might exist. The ease with which we can construct mental models of fictional situations and worlds, is the basis of art and literature. And, of course, one of the biggest contributors to creativity is undoubtedly the social interactions that make up our culture. It is "chance" encounters[10] with others, which provide further opportunities to combine or slightly modify things in new ways, that ultimately drives art, science and society.

20.6 Of Knowledge and Truth

This, of course, leaves us with the question of what knowledge is; what counts as knowledge when anything can be wrong. What we have in our heads are "beliefs"—so the story goes—, and only "justified true beliefs" are real knowledge; false beliefs clearly aren't and true beliefs may be merely accidental—there is no way to "know" whether they are knowledge or not, even if one seems perfectly justified in holding them (the Gettier problem, see (Nagela et al. 2013)).

Peirce suggested that "knowledge", in a fallibilist world, was what scientists would ultimately come to believe "at the limit of enquiry". Whilst clearly correct to some extent, it is not a particularly helpful definition, for even if we appear to have reached the limit of inquiry and scientists have a consensus, they could still conceivably be wrong. History has numerous examples of beliefs that everyone for hundreds of years would have sworn were true, but which ultimately turned out to

[9]The inability to find a suitable analogy has led some physicists to suggest that intuitions provided by realist analogies (models) are unnecessary, and that the highly abstract mathematical formulations (also models) are sufficient in and of themselves. Deutsch disagrees and devotes a large part of his book to explaining his realist model of a quantum mechanical universe.

[10]Modern society promotes individual development by explicitly creating such encounters, in the form of schools, museums, exhibitions, concerts, etc.

be mistaken (or at least not entirely correct), e.g. The Earth is flat, or the Earth is the centre of the universe. Despite Peirce's belief in fallibilism, this view of knowledge and truth still has echoes of inductivism.

A better answer may be the more "pragmatic" (fallibilist) one. Simply put, knowledge in the (old) sense of absolute true facts, does not make sense and must be replaced with a less certain version. This new sense dictates that knowledge is always subject to revision and that there may well be multiple alternative, equally valid, visions of the world. This is not to imply that anything goes; knowledge and truth are still tempered by "reality"—be it the reality of a formal mathematical system, a fictional world set in another time or on another planet, or the "actual" world we all appear to inhabit.

20.7 Concluding Remarks

This paper attempts to understand and respond to Deutsch's recent critique of AI research. It began by trying to clarify some fundamental ideas and terminology related to AI agents. Given this understanding of what agents are and how they may function in the world it became clear that Deutsch makes a valid point when he says that fallibilism is the only basis for creating genuine AI. This is because the choice of philosophical approach affects the ease with which sophisticated agents can create and maintain models. However, when Deutsch claims that AI research has failed to make any progress exactly because it lacked the proper philosophical approach, he is only partially correct, as there are indeed a number of research programs based on fallibilist assumptions. Work by Hutter, Friston, and even Floridi, show both the spark of a new understanding and the difficulty of the undertaking.

Despite calling his approach Universal Induction, Hutter (2005) acknowledges that inducting universal laws is impossible in practice. His elegant theoretical approach to AI thus relies on (what we termed) pseudo-induction, building on work by Smolonsky and Kolmogorov to effectively provide, as he puts it, "a gold standard" for AGI (Rathmanner and Hutter 2011). Unfortunately, like Deutsch's multiverse approach to quantum theory, it involves infinities that make it uncomputable (echoes of inductivism again, perhaps?) In contrast, Andy Clarks' recent survey paper on Predictivism (Clark 2013a, b) describes a more realistic approach to AI, one obviously in line with fallibilism (see also his reply to comments (Clark 2013a, b)). In the paper he examines the work of Friston (2008) and (Friston et al. 2012), whose hierarchical neural network approach to AI sees top-down expectations being compared with input signals to feed-forward error signals (effectively the reverse of most ANN work to date). In another break from tradition, Floridi's (2011) Philosophy of Information offers a semantic view of information as well-formed, meaningful and truthful data. He (mistakenly) views information as being contained wholly "in the message" as it were, rather than in the interpretation of the message by an agent, which is what our analysis would suggest—see also (Adriaans 2013). In this and in his Levels of Abstraction and the Correctness Theory

of Truth, Floridi seems to adopt an absolute observer's perspective, clearly at odds with the fallibilist view presented here. And yet, if one takes a slightly different perspective, as outlined for example in Davenport (2012a, b, 2009, 1997), many of these ideas can align.

Certainly, we are still a long way from developing an AGI, but we do have a better understanding of the problem (and its difficulty). Sadly, a lot of time and research effort is wasted simply because we lack a common vocabulary and approach, surely an indication of a field in a pre-scientific stage. Perhaps Deutsch's "outsider's intuition" can help AI converge on the path to explaining everything.

References

Adriaans, P. (2013). Information. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2013 Edition). Retrieved from http://plato.stanford.edu/archives/fall2013/entries/information/

Bartha, P. (2013). Analogy and analogical reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2013 Edition). Retrieved from http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/

Clark, A. (2013a). Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Sciences, 36*(3), 233–253. doi:10.1017/S0140525X12002440.

Clark, A. (2013b). Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253. doi:10.1017/S0140525X12000477.

Davenport, D. (1997). Towards a computational account of the notion of truth. *TAINN'97.* Retrieved from http://www.cs.bilkent.edu.tr/~david/papers/truth.doc

Davenport, D. (2009). Revisited: A computational account of the notion of truth. In J. Vallverdu (Ed.), *ecap09, Proceedings of the 7th European Conference on Philosophy and Computing.* Barcelona: Universitat Autonoma de Barcelona.

Davenport, D. (2012a). Computationalism: Still the only game in town. *Minds and Machines, 22*(3), 183–190. doi:10.1007/s11023-012-9271-5.

Davenport, D. (2012b). The two (computational) faces of AI. In V. C. Müller (Ed.), *Theory and philosophy of artificial intelligence* (SAPERE). Berlin: Springer.

Deutsch, D. (2011). *The beginning of infinity*. New York: Penguin.

Deutsch, D. (2012, October 3). *Creative blocks the very laws of physics imply that artificial intelligence must be possible. What's holding us up?* Retrieved June 21, 2013, from Aeon Magazine: http://www.aeonmagazine.com/being-human/david-deutsch-artificial-intelligence/

Floridi, L. (2011). *The philosophy of information*. New York: Oxford University Press.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11). doi:10.1371/journal.pcbi.1000211.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151. doi:10.3389/fpsyg.2012.00151.

Goertzel, B. (2012). *The real reasons we don't yet have AGI.* Retrieved from http://www.kurzweilai.net/the-real-reasons-we-dont-have-agi-yet

Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Berlin: Springer. doi:10.1007/b138233.

Kaufman, A. S. (2000). Tests of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence*. New York: Cambridge University Press.

Legg, S. (2008). *Machine super intelligence* (PhD. thesis). Lugano: Faculty of Informatics of the University of Lugano. Retrieved from http://www.vetta.org/documents/Machine_Super_Intelligence.pdf

Nagela, J., Marb, R., & Juan, V. S. (2013). Authentic Gettier cases: a reply to Starmans and Friedman. *Cognition, 129*(3), 666–669.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Rathmanner, S., & Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy, 13*, 1076–1136.

Vickers, J. (2013). The problem of induction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2013 Edition). Retrieved from http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/

Wiedermann, J. (2013). The creativity mechanisms in embodied agents: An explanatory model. *IEEE Symposium Series on Computational Intelligence (SSCI 2013)*. Singapore: IEEE.

Woodward, J. (2011). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2011 Edition). Retrieved from http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/

Chapter 21
Why Emotions Do Not Solve the Frame Problem

Madeleine Ransom

Abstract Attempts to engineer a generally intelligent artificial agent have yet to meet with success, largely due to the (intercontext) frame problem. Given that humans are able to solve this problem on a daily basis, one strategy for making progress in AI is to look for disanalogies between humans and computers that might account for the difference. It has become popular to appeal to the emotions as the means by which the frame problem is solved in human agents. The purpose of this paper is to evaluate the tenability of this proposal, with a primary focus on Dylan Evans' search hypothesis and Antonio Damasio's somatic marker hypothesis. I will argue that while the emotions plausibly help solve the intracontext frame problem, they do not function to solve or help solve the intercontext frame problem, as they are themselves subject to contextual variability.

Keywords Frame problem • Emotions • Search hypothesis • Somatic marker hypothesis • Context

21.1 What Is the Frame Problem?

The frame problem began as an issue in classical artificial intelligence concerning how to represent in formal logic the effects of actions without having to cumbersomely represent the non-effects of actions (McCarthy and Hayes 1969). When an agent acts, the world changes in some ways, but in many others it stays the same. How can a system update its database (or 'beliefs') to reflect these changes? If one simply excludes the non-effects from the program and only represents the effects of the actions then the problem is that it is not a matter of logic that everything else does in fact stay the same. Though it may be a matter of common sense, the artificial system cannot make this deductive inference from the limited information it possesses. If one then opts to include the non-effects in the program, the problem is that this quickly becomes computationally intractable, because the number of non-effects one must include is staggeringly large and leads to a combinatorial explosion.

M. Ransom (✉)
University of British Columbia, Vancouver, BC, Canada
e-mail: madeleineransom@gmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_21

While there are now several adequate methods for addressing this problem in logic-based AI,[1] the frame problem can be expanded into a broader epistemological problem, as Dennett (1978) and Fodor (1983) first noted. Exactly how to characterize this broader problem has remained controversial, however.[2] The issue is that 'the' frame problem is rather a cluster of related problems, though their common core is that each is concerned with how to determine relevance. Wheeler (2008) usefully distinguishes between the *intercontext* and the *intracontext* frame problem. The intracontext frame problem is how to determine, given a context, what information one ought to bring to bear on it – what stored knowledge is relevant in determining what to do? The intracontext frame problem can thus be subdivided into two problems: (i) of the many possible actions, which are relevant and thus deserving of consideration? (ii) of the many possible consequences of these actions, which are relevant? Computational intractability threatens, given the sheer number of possible actions and consequences available for consideration.

Proponents of simple heuristics hold that the intracontext frame problem is solved by human agents via simple, formalizable rules of thumb that bypass complex search procedures and return an answer that is 'good enough' in a wide enough range of situations. The price they pay is optimality – heuristics don't turn up the best solution in all cases. What they gain is speed and computational tractability, and so the intracontext frame problem is disarmed.[3]

While the exact nature of the solution may be disputed, the general form it will take is agreed upon. All that is needed to solve the intracontext frame problem is positing that we humans possess a handy bag of tricks or shortcuts for navigating our way through the vast stores of information at our disposal. However, even if there is a computationally feasible proposal for how we efficiently solve specific types of problems, a more basic and difficult issue remains – how is it that we are able to know which problem we are facing in the first place? The issue is somewhat analogous to the difficulties facing the student of mathematics who has memorized many formulas, but must then recognize not only when a real life situation calls for a mathematical solution, but also which of these formulas to apply.

The intercontext frame problem is precisely that of how to determine what context one is in. It is the problem of determining what features of the environment

[1] See Shanahan (1997). Note, however, that these solutions work only in narrowly fixed domains, and so fail to contribute towards the development of an artificial agent with human scale general intelligence. In the terms I set out below, these proposals can be thought of as solutions to the intracontext, but not the intercontext frame problem.

[2] For an overview of the controversy see Ford and Pylyshyn (1996). Gabbay and Woods (2003, 110–1) identify the frame problem with what AI theorists call the relevance problem.

[3] In another, perhaps compatible approach (see Carruthers 2007), those who hold that the mind has a massively modular architecture may be able to sidestep the intracontext frame problem. These modules are characterized by their dedicated functions – they have more or less narrow purposes and are informationally encapsulated, in that the amount of information the modules are able to draw upon is severely limited. Drastically limiting the amount of information available may render the computational process tractable.

one ought to take as relevant, amongst the many possible candidates, in specifying the situation. What makes the problem so difficult is first that such relevance is largely context-sensitive, and second that we are confronted with ever-changing contexts. For example, whether the fact that there is no milk in the fridge is relevant to an agent depends on the context – is she at the supermarket; does she have money; is she about to go on vacation? Even supposing she is at the supermarket, if an earthquake occurs while she's there, or she realizes that she's late meeting a friend for dinner, then the fact that there's no milk in the fridge ceases to be relevant to the situation. The sheer number of potential contexts dashes any hope of specifying a tractable set of rules for determining relevance in all but the narrowest of domains. Dreyfus (1992) characterizes the problem in terms a 'regress of contexts': the present context can only be recognized in terms of features taken to be relevant in a broader context. This broader context can only be recognized in terms of features taken to be relevant in a still broader context, and so on.[4]

A final point on the scope of the frame problem is that it besieges not only practical but also theoretical reason. Not only are we regularly called upon to come up with a swift answer to the question 'what should I do?' but also to that of 'what should I believe?'[5] The intercontext frame problem here is: what evidence is relevant in determining what to believe? The intracontext frame problem is: given the evidence, what information in one's stored database should one draw upon in order to form one's beliefs? That the frame problem extends to theoretical reason is particularly important when evaluating the role of the emotions, given that many discussions have been confined to the domain of practical reasoning.

21.2 Emotions and the Frame Problem

Attempts to engineer a generally intelligent artificial agent have yet to meet with success, largely due to the (intercontext) frame problem. Given that humans are able to solve this problem on a daily basis, one strategy for making progress in AI

[4]Is the regress infinite? While it may be contexts all the way down, presumably at some point one hits rock bottom. However, while it may not be infinite, the regress is nevertheless still vicious as long as it renders determining context computationally intractable or overly complex and so impractically time-consuming. For example, it would presumably take an extraordinarily long time to determine one's present context if one had to factor in all previous contexts. Stipulating that one only draw upon the relevant factors from previous contexts simply causes the intercontext frame problem to arise anew, as now one must explain how such relevance is determined. Thanks to Chris Mole for discussion on this point.

[5]It is consideration of the second question that has led some philosophers to draw parallels between the problem of induction and the frame problem, though the reduction of the latter to the former is a controversial (and in my view misguided) move. See especially the exchange between Fetzer (1991) and Hayes (1991); and Dennett (1978, 1998).

is to look for disanalogies between humans and computers that might account for the difference. It has become popular to appeal to the emotions as the means by which the frame problem is solved in human agents. While Herbert Simon (1967) pioneered the integration of affect with human cognition, his suggestion that the emotions serve as 'interrupt systems' fell short of addressing the frame problem head on. Ronald de Sousa first picks up the thread (1979) and then makes the explicit connection (1987), arguing that the "Emotions spare us the paralysis potentially introduced by this predicament [of having to first retrieve information in order to determine whether it is relevant] by controlling the salience of features of perception and reasoning" (172).

In what might be construed as a buildup of momentum, several authors have made claims in recent years that emotions (help) solve the frame problem in human, and perhaps artificial, agents. In her landmark book *Affective Computing* (1997), Rosalind Picard argues – while discussing the problem of combinatorial explosion – that "AI has ignored a crucial component [of human intelligence] that is even more basic to human problem solving abilities: the use of feelings and intuition to guide reasoning and decision making [...] An integral component of human decision making is emotion, and this component could potentially be given to computers" (221–2). Dennett (1998) endorses de Sousa's (1979) proposal as a promising avenue for addressing the frame problem, though he laments the lack of a sufficiently concrete scheme for its implementation. Megill and Cogburn (2005) are committed to the hypothesis that "the emotions play a prominent role in preventing humans from suffering from the frame problem" (311). Ketelaar and Todd (2001) argue that "emotions can help the computationally limited human mind to circumvent the pitfalls of the frame problem by determining *which* information to attend to in the first place" (204, emphasis original). Evans (2004) eschews talk of the frame problem altogether in order to avoid controversy over what the problem really is. However, what he calls the 'search problem' – how to cap the number of consequences of an action under consideration – is equivalent to the second part of the intracontext frame problem. Evans takes himself to be elaborating on de Sousa's view, which he characterizes as that the "non-rational procedure for delimiting the range of consequences to be considered in a rational decision process is governed by the emotions" (181).

Damasio's (1994) somatic marker hypothesis (SMH) is often invoked as the means by which the emotions solve or help solve the frame problem. While Damasio himself does not make explicit reference to the frame problem, he holds that the emotions, in the form of somatic markers, "assist the deliberation by highlighting some options (either dangerous or favorable) and eliminating them rapidly from subsequent consideration" (174).

Just how strong the role proposed for the emotions varies, though this is not always made explicit in the literature. This is in part because it is not always clear what the frame problem is taken to be, and thus which problem the emotions are supposed to be solving. Leaving aside the specifics of just how the emotions are

taken to (help) solve the problem in each case, claims can be mapped onto the following taxonomy:

H1: the emotions help solve the intracontext frame problem
H2: the emotions help solve the intercontext frame problem
H1*: the emotions solve the intracontext frame problem
H2*: the emotions solve the intercontext frame problem

One can also be committed to weaker or stronger versions of H1 or H2. Recall that the intracontext frame problem consists of two subdivisions: (i) of the many possible actions, which are relevant and thus deserving of consideration? (ii) of the many possible consequences of these actions, which are relevant? So a weaker version of H1 might hold that the emotions help only with (i) but not (ii), a stronger version that the emotions help with both (i) and (ii). The specific nature and extent of the help the emotions offer will also strengthen or weaken H1 and H2.[6] A further distinction amongst hypotheses that can be made is whether the emotions are taken to (help) solve the (intra/inter-context) frame problem in the domain of practical reasoning only or also in that of theoretical reason. Even if one holds that the emotions alone solve the intracontext frame problem in the domain of practical reason, if this solution does not extend to theoretical reason then the strongest claim one is committed to is H1.

21.3 The Somatic Marker Hypothesis

Damasio (1994) first proposed the SMH as a way of accounting for the behavioral anomalies of numerous patients with damage to the ventromedial prefrontal cortex (VMPFC). While the means-ends reasoning skills of these patients remained unaffected, along with a host of other mental capabilities, they exhibited a striking inability to make wise – or sometimes any – practical decisions regarding their own lives.

The SMH holds that the emotions are involved in practical reasoning – they help us to make decisions concerning what to do.[7] Emotional feelings serve as somatic

[6]Does the strength of the hypotheses also depend on what one counts as an emotion? Perhaps the broader and more inclusive the class or natural kind, the more resources one has at one's disposal for solving the frame problem. On the other hand, increasingly complex and cognitive emotions seem especially subject to the criticisms I make in this paper. However, even if only some subset of the emotions are implicated in solving the frame problem, one could still adhere to H1* or H2*, if these select emotions solve the frame problem on their own. Thanks to Adam Morton for the question, and see Morton (2013) for an example of an inclusive view of the emotions.

[7]Linquist and Bartol (2013) make a distinction between the Somatic Marker Hypothesis and the Somatic Marker Model: "The somatic marker model . . . describes a putative neuro-cognitive mechanism for associating autonomic tags with mental representations. Somatic marker hypotheses, in contrast, invoke this model to explain some aspect of cognition, such as practical decision making"

markers. They 'mark' the content of mental states with either a positive or negative emotional valence.[8] Such markers may then be reconstituted by the agent during deliberation, and thus help guide behavior.

However, as Linquist and Bartol (2013) have pointed out, the SMH isn't just one, but actually a series of separable hypotheses concerning how the emotions are involved in practical reasoning. There are at least five conceptually separable stages in the decision making process where somatic markers may play a role.[9]

In the first stage, decision point recognition, a reconstituted somatic marker functions to alert the agent that a decision should be made. The second stage is that of generating candidate options – somatic markers work to heighten, or 'energize' working memory and attention, allowing the options to be generated. In the third stage, deliberation, the agent identifies the implications or additional properties of various options. The way somatic markers are implicated in the process can be subdivided into two categories. The *relevance hypothesis* is that somatic markers are involved in helping to identity factors relevant to the decision at hand. The *search hypothesis* is that somatic markers are what put a cap on the time and energy we spend deliberating. The fourth stage is that of value assignment and ranking. Value assignment occurs when an option is considered. The somatic marker triggered by the option serves as a factor that weighs for or against it – somatic markers lend valence to the various options. Value ranking is then a means of ordering the various options, thus allowing the option at the top of the hierarchy to emerge as the chosen plan of action. Somatic markers accomplish this by tabulating the valences associated with each option, with the option with the highest overall positive valence winning out. In the fifth stage, somatic markers serve as the motivators for action – once the agent has arrived at a given course of action, somatic markers provide the drive to execute.

21.4 Why Somatic Markers Don't Solve the Frame Problem

21.4.1 The SMH and the Intercontext Frame Problem

The intercontext frame problem arises because there are lots of things we might potentially pay attention to in order to determine context. Decision point recognition thus belongs to the intercontext frame problem – recognizing that a decision is

(458). So, strictly speaking, the SMH may hold that somatic markers are employed in processes other than practical reasoning, such as theoretical reasoning.

[8] Damasio here appears to be using valence to mean what Colombetti (2005) terms 'affect' valence: how good or bad an emotion feels. As she points out, however, the term is used in multiple and sometimes conflicting ways in the literature on the emotions.

[9] To be clear, these are not necessarily sequential stages – many of these may occur in parallel and feed into each other.

called for depends on being able to pick out the features relevant to identifying a (change of) context. The SMH holds that people solve this problem by paying attention to the features of the environment that come with a somatic marker attached. On the stronger reading of this suggestion, corresponding with H2*, people pay attention *only* to those features of the environment that are somatically marked – emotions solve the intercontext frame problem. On the weaker reading, H2, people pay attention *primarily* or *in part* to the valenced features of their surroundings, but this must be supplemented by other strategies or information that guide attention.

The problem with the strong version of the claim – that emotions solve the intercontext frame problem – is that the valence of objects is itself often context dependent. Most objects, people, places, and states of affairs possess both positive and negative aspects, which are often highly variable depending on context. A knife in one context is a helpful tool for making supper and in another it's a threat to one's wellbeing. The prospect of taking a test produces a very different feeling depending on whether one has studied or not. Moreover, there are many concepts which, when combined, elicit an emotional reaction that neither elicits in isolation. As Darwin (1872) observed, a man's beard with some soup caught in it is disgusting even though one typically considers neither soup nor beards to be disgusting. Such contextually elicited emotion is ubiquitous. Which of the multiple somatic markers associated with a given object or situation should be reconstituted in a given situation? Well, the answer goes, it depends on the context. Recognition of which context one is in must therefore come *before* one can employ the appropriate somatic markers, and so they cannot be invoked to solve the intercontext frame problem.

Might the emotions nevertheless *help* solve the intercontext frame problem? Absent a proposal as to what other elements are involved in solving the problem, this claim is difficult to evaluate. However, given that somatic markers themselves depend on context for deployment, the extra component needed to supplement the account appears to be the kind of thing that would itself determine the context, so this other mechanism would be doing all the work. Therefore, the emotions do not solve, nor do they help solve, the intercontext frame problem.

At this point, one might run an objection as follows. Note that H2 itself is ambiguous: it could be either that the emotions solve the intercontext frame problem on their own in a narrow range of cases, or that they are always mere helpers in a wide range of cases; and the first interpretation is not subject to my criticism of H2 here, so it is still a live option.[10] However, then the challenge is how to make the first interpretation tenable – one must explain how it is that the emotions solve the intercontext frame problem on their own in some limited domain. One might proceed by pointing out that while many objects and states of affairs are multi-faceted in their valences, some may be consistently positively or negatively valenced, independent of context. Take the case of fear, for example. When a rabbit

[10]Thanks to Dominic McIver Lopes for this point.

spots a hawk swooping towards it, it just runs away, no matter what activity it was engaged in beforehand. The fear the rabbit feels appears to be what initiates the action, with no need to determine context beforehand. So at least some emotions are capable of guiding action while cutting out the middleman of context.

The first thing to notice here is that while hawks are consistently negatively valenced for rabbits, there are presumably few objects like this for humans. Our world is more nuanced and complex than that of a rabbit, and the vast majority of the objects and situations we encounter will be multi-faceted in their valences. So even if there are a few consistently negatively valenced objects for us, it's hard to see how this will take us any significant distance towards solving the intercontext frame problem. If this is the extent of the help the emotions offer us, then it is exceedingly weak indeed.

Secondly, the action that fear initiates (or perhaps only motivates) in the rabbit does not engage practical reason at all; the action is swift and reflex-like. It's hard to see, therefore, how this sort of action program could be a solution to the intercontext frame problem at all. The frame problem arises on the assumption that we are in fact capable of acting intelligently, where intelligence is defined along the lines of the ability to respond in an adaptive manner to ever-changing contexts (Wheeler 2008), and this is presumed to take place through a rational process of sorts, or at least a flexible cognitive process. While rigid, reflex-like actions may account for our ability to leap out of the way of oncoming threats, for example, they are ill-suited to capture the sort of intelligence we take human beings to possess – the capability of responding to new circumstances with the flexibility needed to navigate them in an advantageous manner. And to respond in such a way, the identification of contextual factors is paramount.

21.4.2 The SMH and the Intracontext Frame Problem

The third and fourth stages of decision-making are where the emotions do the work of addressing the intracontext frame problem. In deliberation, somatic markers flag certain options as relevant, and determine the time to be spent deliberating. In value assignment and ranking, somatic markers serve to further 'prune down' the number of options.

However, for Damasio's proposal to be tenable, he must endorse something along the lines of Newell and Simon's (1976) conception of problem solving as a search through a state space. On this model, rather than generating all the options then pruning them down, individual options are generated and tested step by step.[11] This is precisely the tactic Evans (2004) opts for, and so Damasio's account can

[11]The reason Damasio must endorse this model is because first generating all possible options and then sifting through them simply will not help solve the intracontext frame problem – the sheer number of possible options for any given problem will be enormous, and in some cases infinite.

be supplemented at this point. Evans takes the emotions to solve what he calls the search problem, or the problem of when to stop listing the possible consequences of actions.[12] Invoking Newell and Simon's method, the process of searching for a solution to a given problem can be likened to building a search tree. Potential actions represent the first level of nodes on the tree. Their potential consequences represent the second level of nodes, the consequences of those consequences represent a third level, and so on, with the branches becoming ever denser as one expands the tree. While the tree one can build up is in principle infinite, in practice a good search strategy and test delimit the number of branches to be developed. A search strategy determines which node of the tree ought to be expanded first.[13] Whenever a node is expanded a test is then applied to the result to determine whether it constitutes an acceptable solution. On Evans's view, "emotions prevent us from getting lost in endless explorations of potentially infinite search spaces by providing us both with the right kind of test and the right kind of search strategy for each kind of problem we must solve" (185).

Unfortunately Evans provides no account of how the emotions could be employed to determine which node of the tree to expand first. Perhaps they might work in the form of a rule such as 'expand the most emotionally salient option first.'[14] For such a rule to function, we would need to be able to emotionally appraise the options without considering their consequences, for if we did so we would thereby be expanding the nodes on the tree. This rule would be adequate for cases where the somatic markers are tied directly to the options themselves – perhaps we have already contemplated or experienced their consequences and so have come to mark the option directly.

What test do the emotions provide? It appears that on Evans's view, the emotions function as a test for the viability of a given option by providing valence information. Evans defers back to Damasio at this point – the somatic markers associated with certain consequences count for or against this course of action. If their valence is sufficiently negative, then the option is eliminated. If it's sufficiently positive, then the option is chosen.[15]

[12]The search problem can thus be identified with part (ii) of the intracontext frame problem.

[13]One may, for instance, expand all nodes at the first level before moving on to the next level, or one may choose to expand one particular first-level node on the tree to a fixed depth before moving onto other nodes if the first node provides an unsatisfactory solution.

[14]What if one is indifferent between the options? Then a supplementary rule might added – indifference is a signal to expand all of the first level nodes to the next level, thus taking their consequences into consideration.

[15]What counts as sufficient? While Evans doesn't provide an account, one may suppose there is some threshold, contextually determined by the importance of the situation. Problems deemed extremely important will perhaps require a higher positive score for a given option to be chosen, or a lower negative score for an option to be eliminated. The emotions may be further implicated here, as the strength of one's feelings may serve as a proxy for the subjective importance of the problem, and so serve to set the threshold itself.

However, while the emotions may help to solve the intracontext frame problem, they cannot solve it on their own. The first issue is that in many cases the constraints on what options are generated for consideration are not wholly determined by the emotions. Recall that on Damasio's proposal, somatic markers serve only to sustain the option generation process – they are not implicated at all in the formation of the options themselves. Even supposing they are involved, environmental constraints such as the resources one has at one's disposal, along with background knowledge, habit, and other factors will also go into constraining the types of options that are generated in the first place. For example, taking a private helicopter to work is not an option for most of us – not for lack of enjoyment, but rather due to lack of resources. It is for this reason that the strong claim doesn't go through. The emotions cannot solve the intracontext frame problem on their own because they cannot be wholly responsible for constraining what *sorts* of options are generated in the first place on their own. So the first component of the intracontext frame problem – how to select only the relevant possible actions for consideration – is not wholly resolved by the emotions.[16]

The second issue is that not all cases of practical reasoning are likely to involve somatic markers in any significant way. While there are many emotionally charged decisions we must make in life, so too are there many that will leave us cold. These are not cases where we are indifferent as to the outcome, but rather cases where there aren't enough somatic markers associated with the consequences in order to make this method of option elimination or selection useful. Perhaps there is as of yet no somatic marker associated with the relevant consequences. Perhaps the decision simply calls upon background knowledge rather than somatic markers. This is the second reason why the strong claim is untenable – the emotions can at most help solve the intracontext frame problem because they can't do all the work of eliminating or selecting options in all cases.

Further support for this second point comes from the research of Damasio and colleagues on patients with damage to the ventromedial prefrontal cortex. The evidence suggests that the emotions only play a role in reasoning in situations that directly involve the agent.[17] In the lab, such patients successfully navigate the many possible options to generate reasonable solutions to moral, social or instrumental problems. Given that such patients are hypothesized not to be able to reconstitute their somatic markers when reasoning, such markers cannot be necessary for successful performance on these types of tasks. Therefore, the strongest tenable hypothesis is that somatic markers serve to *help* solve the intracontext frame problem.

A third issue is that the emotions are unlikely to be of much use in solving problems of theoretical reasoning. If one accepts that theoretical reasoning also runs up against the intracontext frame problem, then a proposal about how somatic

[16]Given that Evans's search hypothesis addresses only the second component of the intracontext frame problem, it's unclear that he means to propose a solution to this first issue.

[17]See Damasio (1994), part I.

markers might be of use here is in order. It's hard to see how one might deliver this, though, especially given that the VMPFC-damaged patients show no theoretical reasoning deficits.[18]

A final and more general problem with invoking the emotions is that there is not always a straightforward connection between avoidance and an object or state of affairs that is marked as negative. We have all carried out actions we know we ought to, in spite of not 'feeling' like it. In addition, we actively seek out some fear-producing items, such as horror movies and public speaking. We also seek to avoid many items that we consider pleasurable, such as cigarettes and junk food. The way we assign value to states of affairs thus appears more complicated than mere emotion, suggesting perhaps that value assignment and ranking is not accomplished solely by somatic markers. This in turn speaks to the fact that the emotions can only help solve the frame problem.

In its most successful incarnation then, Damasio and Evans's proposal will be quite weak: it will take emotions to be kinds of heuristics that work along with other heuristics to shrink the space of possible options. The emotions, on this account, are just one method among many to cut down on the number of actions and consequences that need to be considered. So while the proposal can be seen as a serious contender for helping to solve the intracontext frame problem, it falls short of solving the problem on its own. However, one positive element that emerges from this account is that it suggests new directions for empirical research, aimed at exploring the emotions as heuristics model.[19]

21.5 Conclusion

While the emotions may initially appear to offer a promising solution to the frame problem, their helpfulness is severely limited. The emotions don't solve or help to solve the intercontext frame problem because the valence associated with many objects and states of affairs is itself context dependent. The emotions cannot usefully direct us towards relevant features of our environment, because they in turn rely on those same relevant features for their deployment.

[18]There has been recent talk of the epistemic emotions serving as heuristic devices, via somatic markers (Hurley et al. 2011). While the workings of the epistemic emotions remains underexplored, I suspect they will be subject to many of the same objections I raise here. An important challenge the proposal faces is to explain how VMPFC-damaged patients manage to perform well on theoretical reasoning tasks. By hypothesis, these patients cannot reconstitute their somatic markers, so it appears the emotions are not necessary to theoretical reasoning. A possibility (thanks to Samantha Matherne) is that the brain-damaged patients do in fact exhibit some sort of limited theoretical reasoning deficit, perhaps in analogical reasoning.

[19]While Ketelaar & Todd are advocates of the heuristics research program, their claim that the emotions solve the frame problem is too ambitious. The proposal for emotions as heuristics here is more modest.

The emotions don't solve the intracontext frame problem on their own for at least four reasons. First, the emotions are not uniquely responsible for selecting only the relevant possible actions for consideration. Second, not all cases of practical reasoning are likely to involve somatic markers in any significant way. Third, they are unlikely to be of much use in solving problems of theoretical reasoning. Fourth, there is not always a straightforward connection between avoidance and an object or state of affairs that is marked as negative. Therefore, the strongest viable claim is the weak hypothesis, H1, that emotions help solve the intracontext frame problem. Given the diminished prospects for resolving the frame problem in human agents via the emotions, it is unlikely that they will be of much use in addressing the problem with respect to generally intelligent artificial agents.[20]

References

Carruthers, P. (2007). Simple heuristics meet massive modularity. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Volume 2: Culture and cognition* (pp. 181–198). Oxford: Oxford University Press.

Colombetti, G. (2005). Appraising valence. *Journal of Consciousness Studies, 12*(8–10), 103–126.

Damasio, A. (1994). *Decartes' error*. New York: Grosset/Putnam.

Darwin, C. (1872/2009). *The expression of the emotions in man and animals*. New York: Penguin Classic.

De Sousa, R. (1979). The rationality of emotions. *Dialogue, 18*(1), 41–63.

Dennett, D. (1978). *Brainstorms*. Cambridge, MA: MIT Press.

Dennett, D. (1998). Cognitive wheels: The frame problem of AI. In *Brainchildren* (pp. 181–205). New York: Penguin.

Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.

Evans, D. (2004). The search hypothesis of emotion. In D. Evans & P. Cruse (Eds.), *Emotion, evolution, and rationality* (pp. 179–192). Oxford: Oxford University Press.

Fetzer, J. H. (1991). The frame problem: Artificial intelligence meets David Hume. In K. M. Ford & P. J. Hayes (Eds.), *Reasoning agents in a dynamic world: The frame problem*. Oxford: JAI Press.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Ford, K. M., & Pylyshyn, Z. W. (Eds.). (1996). *The Robot's dilemma revisited: The frame problem in artificial intelligence*. Norwood: Ablex.

Gabbay, D., & Woods, J. (2003). *Agenda relevance: A study in formal pragmatics*. New York: North-Holland.

Hayes, P. J. (1991). Artificial intelligence meets David Hume: A reply to Fetzer. In K. M. Ford & P. J. Hayes (Eds.), *Reasoning agents in a dynamic world: The frame problem*. Oxford: JAI Press.

Hurley, M., Dennett, D., & Adams, R. (2011). *Inside jokes: Using humor to reverse engineer the mind*. Cambridge, MA: MIT Press.

[20]For helpful discussion and comments on earlier drafts of this paper thanks to Chris Mole, John Woods, the faculty and graduate students of the University of British Columbia, and the audience of the 2013 Philosophy and Theory of AI conference at Oxford – especially Michael Wheeler and Murray Shanahan.

Ketelaar, T., & Todd, P. M. (2001). Framing our thoughts: Ecological rationality as evolutionary psychology's answer to the frame problem. In P. Davies & H. R. Holcomb (Eds.), *Conceptual challenges in evolutionary psychology: Innovative research strategies* (pp. 179–211). Dordrecht: Kluwer Publishers.

Linquist, S., & Bartol, J. (2013). Two myths about somatic markers. *The British Journal for the Philosophy of Science, 64*(3), 455–484.

McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie & B. Meltzer (Eds.), *Machine intelligence 4* (pp. 463–504). Edinburgh: Edinburgh University Press.

Megill, J. L., & Cogburn, J. (2005). Easy's Gettin' harder all the time: The computational theory and affective states. *Ratio, 18*(3), 306–316.

Morton, A. (2013). *Emotion and imagination*. Cambridge, MA: Polity Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery, 19*, 113–126.

Shanahan, M. P. (1997). *Solving the frame problem: A mathematical investigation of the common sense law of inertia*. Cambridge, MA: MIT Press.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74*(1), 29–39.

Wheeler, M. (2008). Cognition in context: Phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies, 16*(3), 323–349.

Chapter 22
HeX and the Single Anthill: Playing Games with Aunt Hillary

J.M. Bishop, S.J. Nasuto, T. Tanay, E.B. Roesch, and M.C. Spencer

Abstract In a reflective and richly entertaining piece from 1979, Doug Hofstadter playfully imagined a conversation between 'Achilles' and an anthill (the eponymous 'Aunt Hillary'), in which he famously explored many ideas and themes related to cognition and consciousness. For Hofstadter, the anthill is able to carry on a conversation because the ants that compose it play roughly the same role that neurons play in human languaging; unfortunately, Hofstadter's work is notably short on detail suggesting how this magic might be achieved.[1] Conversely in this paper – finally reifying Hofstadter's imagination – we demonstrate how populations of simple ant-like creatures can be organised to solve complex problems; problems that involve the use of forward planning and strategy. Specifically we will demonstrate that populations of such creatures can be configured to play a strategically strong – though tactically weak – game of HeX (a complex strategic game). We subsequently demonstrate how tactical play can be improved by introducing a form of forward planning instantiated via multiple populations of agents; a technique that can be compared to the dynamics of interacting populations of social insects via the concept of *meta-population*. In this way although, *pace* Hofstadter, we do not establish that a meta-population of ants could actually hold a conversation with Achilles, we do successfully introduce Aunt Hillary to the complex, seductive charms of HeX.

Keywords Douglas Hofstadter • Consciousness • Meta-population • Emergence • Swarm intelligence • Stochastic diffusion search

[1] As Drew McDermott writes in the Cambridge Handbook of Consciousness (McDermott 2007), it is as if Hofstadter "*wants to invent a new, playful style of argumentation, in which concepts are broken up and tossed together into so many configurations that the original question one might have asked get shunted aside*".

J.M. Bishop (✉) • T. Tanay
Goldsmiths, University of London, London, UK
e-mail: m.bishop@gold.ac.uk

S.J. Nasuto • E.B. Roesch • M.C. Spencer
University of Reading, UK

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_22

22.1 Swarm Intelligence

In recent years, studies of the behaviour of social insects (e.g. ants and bees) and social animals (e.g. birds and fish) have proposed several new metaheuristics for use in collective intelligence. Natural examples of such 'swarm intelligence' – whereupon apparently intelligent behaviour is realised via various forms of social interaction – include fish schooling, birds flocking, ant colonies in nesting and foraging, bacterial growth, animal herding, brood sorting etc.

Communication – social interaction or information exchange – as observed in social insects is important in all forms of swarm intelligence. In the study of interaction in social insects, two key elements are the individuals and the environment, which results in two modes of interaction: the first defines the way in which individuals interact with each other and the second defines the interaction of individuals with the environment (Bonabeau et al. 2000). Interaction between individual agents is typically carried out via agent recruitment processes and it has been demonstrated that various recruitment strategies are deployed by ants (Holldobler and Wilson 1990) and honey bees (Goodman and Fisher 1991; Seeley 1995). These recruitment strategies may be used, for example, to attract other members of the population to gather around one or more desired areas in the search space, either for foraging purposes or in order to facilitate a colony relocation to a better nest site.

It has been observed that recruitment strategies in social insects may take several forms: localised or global recruitment; one-to-one or one-to-many recruitment; and may operate stochastically or deterministically. The nature of information exchange also varies in different environments and with different types of social insects. Sometimes the information exchange is quite complex and, for example, might communicate data about the direction, distance and suitability of the target; or sometimes the information sharing is relatively simple, for example, a stimulation forcing a particular triggered action. Nonetheless, what all recruitment and information exchange strategies have in common is an ability to distribute useful information across their community (De Meyer et al. 2006).

Chemical communication through pheromones forms the primary method of recruitment in many species of ants, however in one species, *Leptothorax acervorum*, a 'tandem calling' mechanism (one-to-one communication) is used. In this process, on its return to the nest, a forager ant that has found the resource location physically recruits a single ant and, by this action, the location of the resource is physically publicised (Moglich et al. 1974) to the population.

Swarm intelligence, as the study of metaheuristics inspired by natural collective intelligence, is a relatively new branch of artificial intelligence that realigns intelligence away from the individual towards the collective; its aim is to illustrate intelligent behaviour by considering individuals in a social context and monitoring their interaction with one another as well as with their environment. Natural examples of swarm intelligence systems are: fish-schooling, bird-flocking, animal herding, nesting and foraging in the social insects etc. In recent years, abstractions

of such natural behaviour have motivated several new Swarm Intelligence heuristics. While in typical Swarm Intelligence algorithms only the syntactical exchange of information is considered, in many natural social interactions, it is not just syntactical information, but also semantic rules and beliefs about how to process this information, that is exchanged (Kennedy et al. 2001).

The simple and often successful deployment of swarm based heuristics on traditionally difficult optimisation problems has generated significant interest (cf. Dorigo et al. 1991; Dorigo 1992; Kennedy and Eberhart 1995); nonetheless, to date, they have merely been deployed on conceptually straightforward optimisation and regression problems.

This paper is organised in the following manner. In Sect. 22.2 we introduce the game of Hex. In Sect. 22.3 we introduce a Monte Carlo Stochastic Diffusion Search (MCSDS), a swarm intelligence algorithm for playing Hex based on a simple merger of Stochastic Diffusion Search (SDS) (Bishop 1989) and Monte Carlo methods (Metropolis and Ulam 1949). Subsequently extending MCSDS in Sect. 22.4, we introduce a more sophisticated algorithm, Stochastic Diffusion Search applied to Trees (SDST); a novel swarm intelligence heuristic able to solve the complex and general problem of forward planning in a way analogous to Monte-Carlo Tree Search (MCTS) (Abramson 1990).

In SDS and MCSDS, direct one-to-one communication (which is similar to the tandem calling recruitment mechanism described earlier) is utilised.[2] In SDST, each individual agent processes information concerning a unique action without "awareness" of the way in which actions are being compared and combined. Yet the dynamics of the entire population of agents lead to a high level "reasoning" about successions of actions analogous to Monte-Carlo Tree Search (MCTS). In its functioning, SDST is argued to introduce a meta-level in the swarm intelligence paradigm.

Although some previous attempts have been made to apply decentralised methods to forward planning tasks, such methods did not reach the same degree of generality as SDST. For example Tesauro developed in 1989 a neural network program playing Backgammon (a non-deterministic finite two-person zero-sum game with perfect information) better than any other program (the program called Neurogammon won the backgammon competition of the First Computer Olympiad). However, Tesauro explicitly expressed in the introduction of Hart (1992) that "the game of backgammon in particular was selected because of the predominance of judgement based on static pattern recognition, as opposed to explicit look-ahead or tree-search computations."

By presenting SDST, our objective is to extend the applicability of parallel and distributed models of computation (and in particular SDS) to solve problems that were historically exclusively addressed with a sequential algorithmic approach requiring centralised control and access to the data. For the sake of simplicity and

[2] Although the recruitment behaviour of real ants is more complex than the behaviour in SDS, both are population-based and find their optima via agents communicating with each other.

clarity, and because it is the problem for which it was originally conceived, SDST is presented in the context of combinatorial games (finite two-person zero-sum games with perfect information such as Chess). However, the discussion is entirely consistent with any planning task that can be represented as a tree of sequential decisions. Along the way we will illustrate how Hofstadter's 'Aunt Hillary' might beat her old friend the 'Ant Eater' at Hex.

22.2 The Game of Hex

Hex is a combinatorial game that belongs to the family of connection games. It is an example of a game that has been solved in a non-constructive way: it is proved that the first player has a winning strategy, but the strategy in question is not known. The proof of this result is given in the following sections, after the rules of the game have been presented. The game of Hex was chosen for three main reasons:

- First, it is relatively simple and well-suited to perform rapid random game evaluations: the only action available to each player at each turn is to put one stone of his colour on the board.
- Second, the size of the board is changeable without modifying the nature of the game. This allows the algorithm to be tested on different sizes of game-trees.
- Third, it is mathematically elegant and has interesting properties in game theory.

22.2.1 Rules and History

The game of Hex was first invented in 1942 by the Danish scientist, artist and poet Piet Hein. It was then independently reinvented in 1947 by the American mathematician John Nash, while still a student at Princeton University. It was first known under the name *Polygon* in Denmark and was called after its creator *John* or *Nash* at Princeton University, before Parker Brothers marketed a version under the name *Hex* in 1952.

The game is played on a rhombic board covered with hexagonal cells. Each of the two players is associated with a colour (blue and red in the following) and two opposite sides of the board (for example top-right and bottom-left for blue and top-left and bottom-right for red). The rules are extremely simple: the two players alternatively place a stone of their colour on a single cell within the entire board and try to form a connected path between their two sides. The game ends when one of the two players managed to build such a connection. The usual size of the board is 11×11, but due to the relationship that Hex maintains with Go, the sizes 13×13 and 19×19 are also common. According to Sylvia Nasar's biography of John Nash *A Beautiful Mind*, he recommended 14×14 as the optimal size. Figure 22.1 shows a typical finished game: Red wins because he managed to connect his two sides of the board.

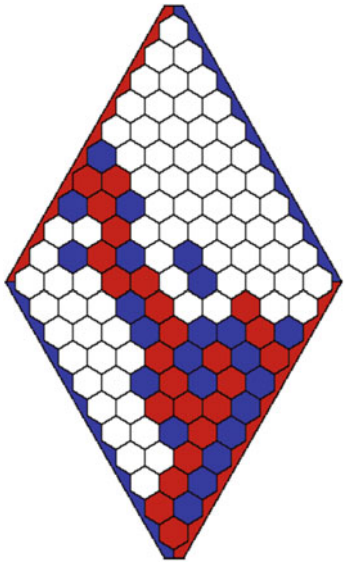


Fig. 22.1 A typical finished game at Hex (From Wikimedia Commons)

22.2.2 Game Theory

The proof of the existence of a winning strategy for the first player relies on two central points: the fact that there can be no draw in Hex, and the strategy stealing argument. A relatively simple proof of the first point is given in a paper from David Gale: "The game of Hex and the Brouwer fixed-point theorem" and is outlined in the next section. Interestingly, the same paper establishes an equivalence between this proof and the Brouwer fixed-point theorem, an important theorem in topology that states that for any continuous function f with certain properties there is a point x_0 such that $f(x_0) = x_0$.

22.2.2.1 One and Only One Winner

If the blue player is called x and the red player is called o, and if the two sides of the board corresponding to the blue player are called X and X' and the two sides corresponding to the red player are called O and O', Gale states what he calls the Hex theorem as follow:

Hex theorem: If every tile of the Hex board is marked either x or o, then there is either an x-path connecting regions X and X' or an o-path connecting regions O and O', but not both. (Gale 1979)

In the original paper, Gale gives a very intuitive illustration of the theorem: "Imagine, for example, that the X-regions are portions of opposite banks of the river "O" (...) and that the x-player is trying to build a dam by putting down

stones. It is quite clear that he will have succeeded in damming the river only if he has placed his stones in a way which enables him to walk on them from one bank to the other." In other words, the only way one has to prevent his opponent from winning is by winning himself and thus there is always a winner. Then Gale continues his analogy: "if the x-player succeeds in constructing a causeway from X to X', he will in the process have dammed the river and prevented any flow from O to O'." In other words, if one of the players wins he prevents his opponent from winning at the same time and there can be only one winner. Although the theorem is intuitive, the proofs of the two results (the existence and the uniqueness of a winner) are rather delicate. The uniqueness directly follows from a fundamental result of topology called the Jordan Curve Theorem. This theorem asserts that every non-self-intersecting continuous loop divides the plane in an "interior" region and an "exterior" region so that any continuous path connecting a point of one region to a point of the other intersects with that loop somewhere. Although this theorem is also very intuitive, it is difficult to establish formally.

A constructive proof of the existence of a winner is given in Gale (1979). It is outlined in the first part of the paper as follows (Fig. 22.2 is the corresponding figure on which colours have been added for clarity):

> We consider the edge graph Γ of the Hex board to which additional edges ending in vertices u, u', v, v' have been added to separate the four boundary regions, as shown in the figure. We now present an algorithm for finding a winning set on the completely marked board. We shall make a tour along Γ, starting from the vertex u and following the simple rule of always proceeding along an edge which is the common boundary of an X-face and an O-face. Note that the edge from u has this property since it separates regions X and O. The key observation is that this touring rule determines a unique path; for suppose one has proceeded along some edge e and arrives at a vertex w. Two of the three faces incident to w are those of which e is the common boundary, hence one is an X-face, the other an O-face. The third face incident to w may be either an X-face or an O-face, but in either case there is exactly one edge e' which satisfies the touring rule.

The tour constituted by this algorithm is highlighted in yellow in Fig. 22.2 (it starts at vertex u at the bottom and ends at vertex v on the left). Two characteristics of this tour show that there necessarily is a winner:

1. It will never revisit any vertex. The reason for this is that by construction, the degree of every vertex of the tour is at most two. But since the degree of the first vertex of the tour is one (vertex u), the tour must constitute a simple path and end on a vertex of degree one. The only possibilities are u', v or v'.
2. It cannot end on vertex u' (top one). The reason of this is that the tour starting at u always keep blue cells on the left and red cells on the right while the edge ending at u has a red cell on its left and a blue cell on its right.[3] Hence the tour can only end on v or v'.

[3]The rigorous proof is not based on this left-right consideration: "this would involve getting into the quite complex notion of orientation, which is not needed for our proof" (Gale 1979).

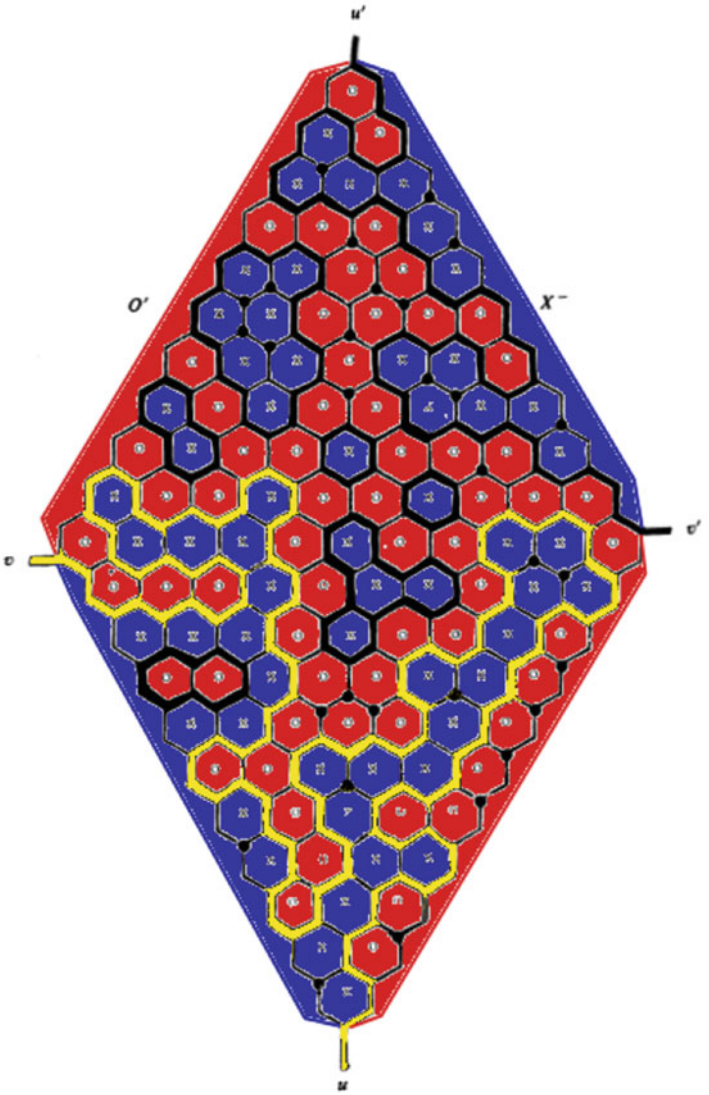


Fig. 22.2 Illustration of the tour followed by Gale's algorithm on a full Hex board (starting at u and finishing at v in *yellow*). The *red cells* on its *right* form a winning path for o (From Gale 1979, colours added)

To conclude, one only needs to notice that if the tour ends on v the red cells on its right form a winning path for o and if it ends on v' the blue cells on its left form a winning path for x. Hence, there is always a winner at Hex.

22.2.2.2 The Strategy Stealing Argument

The absence of draw at Hex has an interesting consequence if one remembers Zermelo's theorem: according to Hart (1992), Zermelo's theorem states that "in Chess, either White can force a win, or Black can force a win, or both sides can force a draw." Now it can be maintained that in Hex, either Blue can force a win or Red can force a win. In fact, an ingenious reductio ad absurdum from John Nash called the strategy stealing argument proves that this is the first player who has a winning strategy. The argument goes as follow: let one suppose that there exists a winning strategy for the second player. In this case, the first player can steal this strategy to build his own winning strategy in the following way. First he places one of his stones anywhere on the board and let the second player play as if he was the first one. Then he follows the second player's winning strategy until either the game finishes, or the winning strategy tells him to play the move he played first. In the second case, he just plays anywhere and starts following the winning strategy again the next turn. In this situation both players have a winning strategy which is contradictory and the initial hypothesis that the second player has a winning strategy is false. It is important to note here that the stealing strategy argument only holds because it is never a disadvantage to play a move at Hex. However this is not always the case: in Chess there are situations called *Zugzwang* in which every move leads to a worse and often lost position (they happen most of the time in late endgames).

Of course, the strategy stealing argument is non-constructive and the winning strategy for the first player is not known for boards bigger than 9×9 cells. Yet, in practical play, it appears that playing first does constitute a great advantage. To compensate this bias, the swap rule allows the second player to choose between either playing normally, or taking the first player's position after his first move. This re-equilibrates the game because in this case the first player should play neither the strongest moves (such as the centre of the board) because the second player would switch its position with him, nor the worst moves (such as the two cells in the acute angles of the rhombus) because the second player would leave them to him. In the presence of the swap rule this is the second player who has a winning strategy since he can choose between taking the first player's move if it is a winning one or leaving it if it is a loosing one (although this information is not known in practical play).

22.3 Playing Games with Aunt Hillary

The work presented in this section rests on two pillars: a swarm intelligence metaheuristic for search and optimisation called Stochastic Diffusion Search (SDS) and the Monte-Carlo search method. These two techniques are briefly outlined in the following subsections.

22.3.1 Stochastic Diffusion Search (SDS)

SDS is an efficient probabilistic swarm intelligence global search and optimisation technique that has been applied to diverse problems such as site selection for wireless networks (Whitaker and Hurley 2002), mobile robot self-localisation (Beattie and Bishop 1998), object recognition (Bishop 1992) and text search (Bishop 1989). Additionally, a hybrid SDS and n-tuple RAM (Aleksander and Stonham 1979) technique has been used to track facial features in video sequences (Bishop 1992; Grech-Cini and McKee 1993). Previous analysis of SDS has investigated its global convergence (Nasuto and Bishop 1999), linear time complexity (Nasuto et al. 1998) and resource allocation (Nasuto 1999) under a variety of search conditions.

SDS is based on distributed computation, in which the operations of simple computational units, or agents are inherently probabilistic. Agents collectively construct the solution by performing independent searches followed by diffusion of information through the population. SDS relies on two principles: partial evaluation of hypotheses and direct communication between agents. The SDS algorithm is characterised by three phases: Initialisation, Test and Diffusion – the test and diffusion phases are repeated until a Halting criterion is reached. During the initialisation phase each agent formulates a hypothesis, i.e. chooses a potential solution in the search space. During the test phase each agent partially evaluates its hypothesis: agents for which the partial evaluation is positive become active, and the others become inactive. During the diffusion phase, agents exchange information by direct communication: each inactive agent X contacts an agent Y at random. If Y is active, X takes its hypothesis, otherwise X formulates a new hypothesis at random (procedure called passive recruitment). In practice, a halting criterion needs to be defined to stop the algorithm running; the properties of convergence of SDS led to the definition of two criteria, a weak and a strong version (Nasuto and Bishop 1999).

22.3.2 Monte-Carlo Stochastic Diffusion Search (MCSDS)

The starting point in applying Monte Carlo methods to SDS is the simulation of random games a great number of times. The suitability of Hex to perform random game simulations was one of the main reasons to select this game as a study case. Indeed, a random game just consists in alternatively placing red and blue stones on the board until it is full. Although this is relatively simple, some care has to be given to fill the board with a uniform distribution or the evaluation of the moves could be biased.

A simple improvement that can be given to the standard Monte-Carlo algorithm is to sample the evaluation of the cells by applying a ‘multi-armed bandit’ analogy’: the moves that tend to give good results at the beginning of the evaluation should receive more attention than the moves that appear to be bad. This can be achieved by applying SDS.

First, a population of agents with hypotheses about the best move to play is initialised. Second, the hypotheses are tested by performing a random game simulation: if the outcome is a win the agent becomes active, otherwise stays inactive. Third, every inactive agent selects at random another agent for communication. If the selected agent is active, the first agent copies its hypothesis, but if the selected agent is inactive, the first agent chooses a new hypothesis at random (passive recruitment strategy).

It is important to notice that this 'improvement' only concerns the speed of the process; theoretically the value attributed to each move is the same as in the standard version. Indeed, the value of a move is still assimilated to the probability that it leads to a win given random play, and this probability is not changed by the way evaluation is balanced between the different moves. Hence using SDS in this way has exactly the same 'level of play' as standard Monte-Carlo – it simply uses SDS as an efficient Swarm Intelligence resource allocation technique.

22.3.3 Hex and the Single Ant Hill

Hofstadter (1979) imagines Ant Hillary and the Ant Eater conversing as follows:

> ANTEATER: ... Aunt Hillary and I have conversations for hours. I take a stick and draw trails in the moist ground, and watch the ants follow my trails. Presently, a new trail starts getting formed somewhere. I greatly enjoy watching trails develop. As they are forming, I anticipate how they will continue (and more often I am wrong than right). When the trail is completed, I know what Aunt Hillary is thinking, and I in turn make my reply.

In this paper we do not, *pace* Hofstadter, illustrate Aunt Hillary in communication with the Ant Eater; we do however, successfully introduce Aunt Hillary (AH) to the complex, seductive charms of HeX. To do this we simply need to show that Aunt Hillary can perform the Monte-Carlo SDS Hex algorithm outlined in Sect. 22.3.2.

In nature, ants produce numerous different pheromones, each with its own distinct purpose. For example, ants secrete pheromones to attract mates, to signal danger to the colony, or to give directions about a location. Ant Hillary deploys pheromones to represent information and, in this way, can play a game of Hex against the Ant Eater (AE) as follows.

Aunt Hillary maintains a group of red ants to identify hypotheses about the potential moves available at the current stage of the game; a *partial evaluation* of each move is constituted by the result of a random game performed assuming a hypothesized move has been played; after which, SDS diffuses 'successful' hypotheses through the population. After a number of these hypothetical plays have been evaluated, red ants will tend to cluster around moves that have the highest probabilities of leading to a win. This process constitutes the simplest application of MCSDS to the problem of computer game-playing and is easily implemented as a simple 'ant algorithm' that Aunt Hillary can physically enact.

Aunt Hillary first divides the ants she will deploy in the game into three types, defining *hypothetical-moves, random-plays and actual-moves*:

- a population of k red 'H-ants' – each of which maintain a *Hypothesis* suggesting Aunt Hillary's next best move onto an unoccupied position on the board; each H-ant carries a unique pheromone representing that H-ant's hypothesis;
- associated with each H-ant are two equal sized groups of g[4] black and white 'R-ants'; each R-ant representing a potential legal *Random* play from either Aunt Hillary [black] or Ant Eater [white]. Each R-ant is uniquely associated with a particular hypothesis by carrying that H-ant's pheromone. Once a 'hypothetical move' has been made by Aunt Hilary, successive deployment of associated (black and white) R-ants thus demarcate positions on the board with successive *random moves* played by the Ant Eater and Aunt Hillary;
- two equal sized groups of g black and white 'M-ants'; each M-ant representing an actual played *Move* from either Aunt Hillary [black] or Ant Eater [white]. Black and white M-ants thus demarcate positions on the board with the successive *actual moves* played by either Aunt Hillary or the Ant Eater.

Aunt Hillary initially allocates the red H-ants with a random hypothesis, i.e. a hypothetical 'next-play' selected randomly from the current set of legal moves available on the board. When it is her turn to move, Aunt Hillary 'thinks' until her thinking time is up, after which, she moves to the position defined by the most popular hypothesis (i.e. the hypothesis carried by most red H-ants).

Aunt Hillary subsequently makes this move by positioning a black M-ant to this board position. The Ant Eater then makes its play and AH marks its move by positioning a white M-ant to demarcate that board location. After a finite sequence of such turns – as Nash proved – either Aunt Hillary will have won (established a continuous path [linking AH's M-ants] from left to right across the board) or the Ant Eater will have won (established a continuous path [linking AE's M-ants] from the top to the bottom of the board).

In performing her 'thinking' Aunt Hillary merely iterates the following two-step parallel procedure until her 'thinking time' is over and a move is obliged.

1. *Evaluate Monte-Carlo games*. For each of the [red H-ant] hypotheses Aunt Hillary performs a Monte-Carlo simulation and plays a 'random game'. To do this, the population of black and white R-ants [associated with each focal H-ant hypothesis] take turns to play randomly selected legal moves – whereby a (black/white) R-ant walks to a random position (unoccupied by either an R-ant associated [with this hypothesis] or a previously played M-ant) until either Aunt Hillary has won (established a continuous path [linking AH M-ants, the focal H-ant and black R-ants associated with the focal ant's hypothesis] from left to right across the board) or the Ant Eater has won (established a continuous path [linking AE M-ants and the white R-ants associated with each focal hypothesis] from the top to the bottom of the board). If the random game resulted in a win for the Ant Eater then the focal H-ant hypothesis is deemed *inactive*; if the result was a win for Aunt Hillary the focal hypothesis is *active*.

[4]Group size $= g =$ (board area DIV 2) $+ 1$.

In performing each random move, a R-ant randomly walks around the board for a random time period (t) after which it stops at the first position not occupied by either an ant baring the same hypothesis pheromone or a M-ant (demarcating a previously played move).

2. *Diffuse hypotheses*. Each *inactive* red H-ant randomly selects another H-ant; if that H-ant is active, it transfers its hypothesis (active H-ant → inactive H-ant) otherwise the inactive H-ant selects a new hypothesis at random from the unoccupied board positions. To physically perform such 'hypothesis diffusion', each H-ant [hypothesis] merely moves randomly around the board for a random time period (t) after which it continues to move randomly until it alights onto a position occupied by another red H-ant. If this H-ant is *active*, it stays at this position; if it is *inactive*, it continues to move randomly stopping as soon as it alights onto a position not occupied by a previous move (as demarcated by the presence of an M-ant).[5] In other words, if diffusion did not occur the H-ant selects a new hypothesis at random by simply randomly moving around the board for a random time period (t), after which it stops at the first position not occupied by a M-ant.

NB. As the pheromones carried of each of the H-ants (and their associated group of R-ants) are distinct, the Monte-Carlo games for all of the hypotheses and diffusions processes can be evaluated in parallel; *thus Aunt Hillary's 'thinking' is characterised by a seething mass of randomly moving ants*.

22.3.4 Analysis

While both the standard Monte-Carlo program and MCSDS show relatively good strategical sense and always perform moves that increase the overall chance to win, both play poor tactically. In early evaluation trials, it was established that the MCSDS algorithm offered good performance on a 7×7 Hex board against a naive random opponent, but poor tactical play against a more skilled opponent.

An illustration of one of the tactical weakness of MCSDS concerns situations called 'bridges'. In Hex, a bridge situation occurs whenever a player cannot stop the other player from connecting two groups of stones in one move because there are two ways to do it (see Fig. 22.3). When a bridge is formed for one player, the best tactic for the other player is to play somewhere other than closing the bridge. However, the MCSDS algorithm (as deployed by Aunt Hillary) is unable to reliably play this way; to do so would require the anticipation of potential next moves from the opponent (see Fig. 22.4).

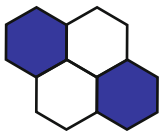


Fig. 22.3 A bridge for *Blue*: *Red* cannot stop *Blue* from connecting the two *blue cells* on the next move

[5] A process isomorphic to asynchronous passive recruitment SDS (De Meyer 2003).

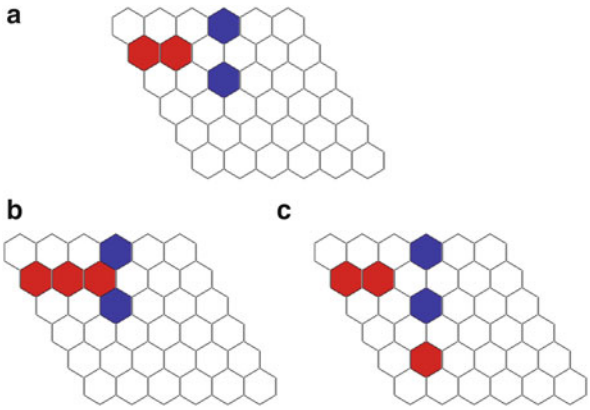


Fig. 22.4 Illustration of the weakness in tactic peculiar to the standard Monte-Carlo program. (**a**) An hypothetical game situation. It is *Red* to move. (**b**) The move that the standard Monte-Carlo program would perform: it increases its chances to win if the rest of the game is played randomly. (**c**) Example of a better tactical move

To improve the tactical play, a metapopulation (Levins 1969) of ants is deployed to enable a Monte-Carlo informed SDS to tactically explore the game tree[6]: Stochastic Diffusion Search applied to Trees (SDST).

22.3.5 Monte-Carlo Tree Search (MCTS)

MCTS "is a recently proposed search method that combines the precision of tree search with the generality of random sampling" (Browne et al. 2012). To date, over 350 papers related to MCTS have been published, with applications ranging from computer Go to Constraints Satisfaction problems through Reinforcement Learning and Combinatorial Optimisation. Browne et al. (2012) offer a complete survey of the published work on MCTS (until 2011) and conclude that MCTS "has already had a profound impact on Artificial Intelligence (AI) approaches for domains that can be represented as trees of sequential decisions, particularly games and planning problems".

MCTS has originally been developed in the context of computer game playing and finds its roots in B. Abramson's 1990 paper *Expected-outcome: a general model of static evaluation* (Abramson 1990). This paper introduces the central Monte Carlo theme to evaluate a game position by playing a large number of random games from that position, assuming that a good move must increase the expected outcome of the player.[7] The second decisive step in the development of MCTS was the publication in 2006 of Kocsis and Szepesvári's paper *Bandit based Monte-Carlo Planning*.

[6] A 'game tree' is a directed graph whose nodes are positions in a game and whose edges are moves. The complete game tree for a game is the game tree starting at the initial position and containing all possible moves from each position; a n-ply game tree describes all possible move/counter-move combinations to a depth of n moves.

[7] This assumption is not necessarily a good one due to the distinction between random play and optimal play – see analysis of standard Monte-Carlo methods (and MCSDS) in Sect. 22.3.4.

In this paper the 'Upper Confidence bound applied to Trees' (UCT) heuristic is introduced; a method that "applies bandit ideas to guide Monte-Carlo planning". The crux of UCT is to choose the moves to be evaluated at each node of the game-tree according to the information already collected during previous evaluations, in order to better exploit the most promising areas of the tree. Standard MCTS consists in iteratively building a "search-tree" (the root node of which is the current position) and is outlined in Chaslot et al. (2008) as a succession of four phases: Selection, Expansion, Simulation and Backpropagation. In practice, the four phases are repeated until a given computational budget is spent (usually the time), at which point a decision is made and a move is played.

The moves to be evaluated are first chosen in the existing search-tree from the root in a way that balances between exploration of the available moves and exploitation of the most promising ones (*selection*): the policy used to choose the moves during this phase is called the "tree policy" and this is where (Kocsis and Szepesvári 2006) introduced the analogy between a node of the search-tree and a multi-armed bandit.

When a leaf of the search-tree is reached, the rest of the game is played up to a final state (*simulation*). The policy used during this phase is called the "default policy" and can be purely random in the simplest implementations of MCTS. The first move chosen by the default policy is then added to the search-tree (*expansion*).

Finally, the statistics of each node crossed during the selection phase are updated according to the outcome of the simulated game (*backpropagation*).

The way MCTS works is rather intuitive and it is argued in Browne et al. (2012) that "the forward sampling approach is, in some ways, similar to the method employed by human game players, as the algorithm will focus on more promising lines of play while occasionally checking apparently weaker options." An important property of MCTS is its asymptotic convergence to Minimax, i.e. it is assured to select the best move available if enough time is given (albeit the convergence to Minimax can take a very long in practice).

22.4 Stochastic Diffusion Search Applied to Trees (SDST)

Conceptually, the application of SDS to game-tree exploration is a two step process. First, each node is attributed a distinct and independent *local population* of agents to solve the problem of move selection on that node. Second, a reallocation policy is used to move the uncontacted agents toward more interesting regions of the game-tree – thus leading to the formation of a dynamically moving *metapopulation*[8] of agents.[9]

[8]The term was coined by Levins in Levins (1969) to describe the dynamics of interacting populations of social insects.

[9]The initial motivation for the work on SDST was to extend the applicability of Stochastic Diffusion Search (SDS) to more complex search spaces, and combinatorial games were chosen

Table 22.1 First application of SDS to game-tree exploration: use of multiple populations of agents

Initialisation	During the initialisation phase, a local population of agents is generated for each node of the game-tree up to a fixed depth D. For each local population, agents' hypotheses are initialised to a possible move of the corresponding node.
Test	During the test phase, a complete hypothesis is formed for each agent in the local population corresponding to the root node (later called root node population). This is done by combining agents from different local populations in a way analogous to the selection phase in MCTS: for each agent X in the root node population, an agent Y in the local population pointed by X's hypothesis is selected. Then an agent in the local population pointed by Y's hypothesis is selected, etc, until depth D is reached. Once a hypothesis is formulated, a simulation is run (in the MCTS sense) and the activities of the agents forming the hypothesis are updated according to the node they belong to (step corresponding to the backpropagation in MCTS): if the simulation leads to a win for Max, the agents in populations corresponding to Max's nodes become active and the agents in populations corresponding to Min's nodes become inactive (if it leads to a loss, it is the contrary).
Diffusion	During the diffusion phase, each local population acts independently, i.e. a diffusion phase is undertaken in the sense of Standard SDS without communication with other local populations.

22.4.1 First Step: Use of Multiple Populations of Agents

The first step toward implementing SDST is to use SDS to solve the "exploration-exploitation dilemma" appearing during the selection phase of MCTS at each node of the search-tree. An algorithm detailing this idea is given in Table 22.1 (in SDS terms).

The operation of this algorithm is illustrated in Fig. 22.6 on the small game-tree presented in Fig. 22.5. The studied game-tree has been specifically designed to reveal the ability of the algorithm to converge to minimax and escape local optima: while a monte-carlo evaluation of the left and right moves for Max at the first ply would respectively lead to 50 % and 75 % chances to win – thus suggesting that the right move is better – the minimax resolution of the game-tree actually shows that, *if the players play optimally*, the left move leads to a win for Max (whatever Min plays at the second ply, the right move for Max at the third ply leads to a win) while the right move leads to a loss (if Min plays his left move at the second ply, whatever Max plays for the third ply leads to a loss with Min playing the left move at the fourth ply).

as a first study case. Then, Monte-Carlo Tree Search (MCTS) came naturally as a good framework for several reasons. First, MCTS does not rely on domain knowledge but rather on a large number random game simulations and the notion of random game simulation fits well with the concept of partial evaluation in SDS. Second, the strength of MCTS relies on the tree policy balancing between exploration of the search space and exploitation of the promising solutions and SDS is a metaheuristic precisely conceived to solve this "exploration-exploitation dilemma" in the management of the computational resources. Finally, MCTS has proven very successful in a wide range of problems – not only game playing – and is still under active study.

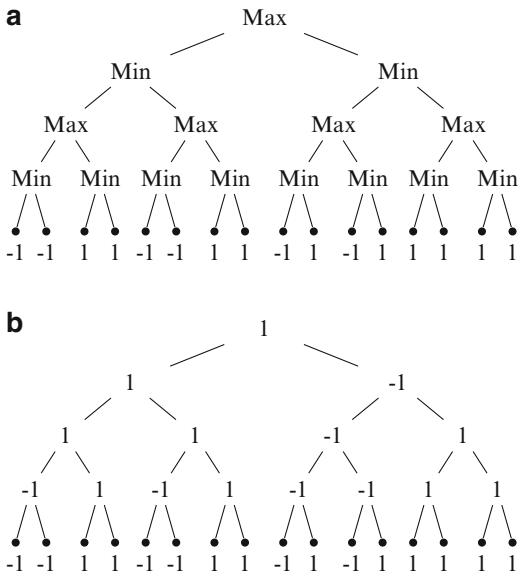


Fig. 22.5 Studied game-Tree. The minimax resolution shows that Max is the winner if he plays optimally. (**a**) Studied game-tree. (**b**) Minimax resolution

Figure 22.6 shows that during iterations 1 and 2, most of the agents in the root node population point toward the right move. Then during iterations 3 and 4, the selection of Min's left moves at plies 2 and 4 changes this tendency and at iteration 5 all the agents in the root node point toward Max's left move – the best move in the minimax sense. Figure 22.6 simply illustrates that, as any other MCTS with a different tree policy, the algorithm presented here converges to minimax (provided that every non-terminal node of the game-tree is being attributed a population of agents).

22.4.2 Second Step: Use of a Reallocation Policy

Although the previously discussed algorithm is shown to solve the problem of game-tree exploration, it suffers from two main drawbacks. First, the number of studied nodes in the game-tree and the number of agents per node need to be fixed manually in a very artificial way. Second, a uniform repartition of the agents in the initialisation phase rapidly leads to many agents being uncontacted in some branches (for example, all the agents on the right side of the tree become useless after the fifth iteration in Fig. 22.6).

These drawbacks can be solved with the use of a reallocation policy where agents are scattered in the tree from the root node and uncontacted agents are backscattered toward parent nodes. SDST uses such a reallocation policy, defined naturally as described in Table 22.2.

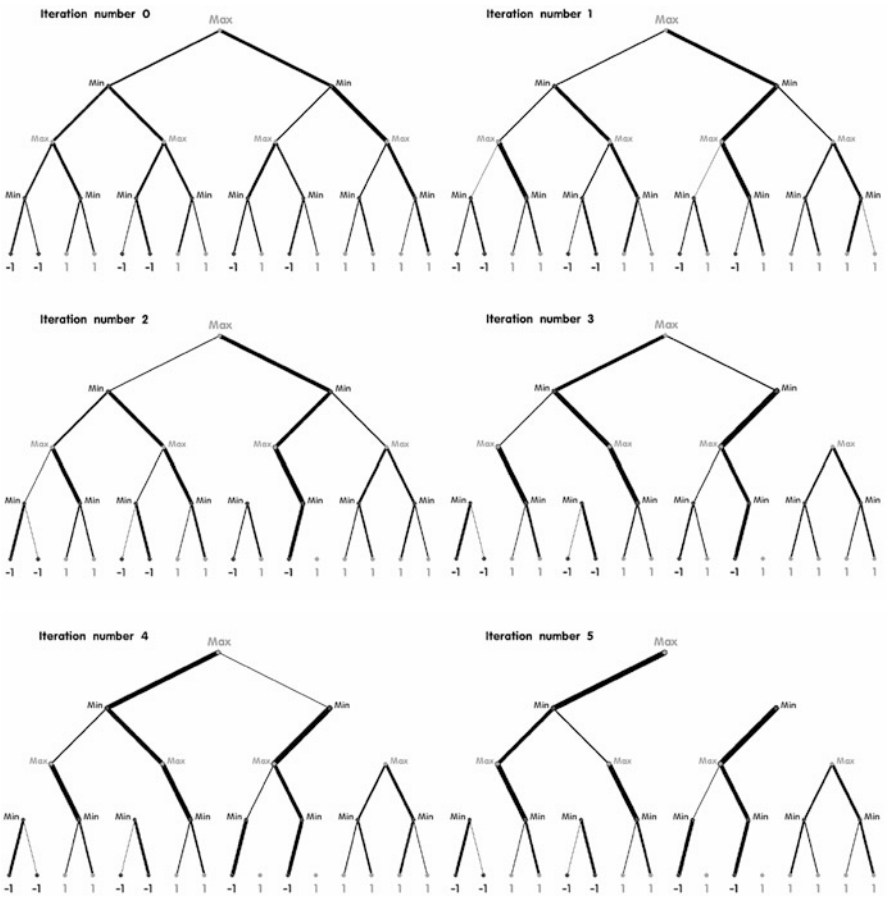


Fig. 22.6 Illustration of the algorithm presented in Table 22.1: Evolution of the distribution of the agents in the different nodes of the studied game-tree (first 5 iterations shown, total number of agents = 175). Each branch has an area proportional to the number of agents in the parent node population supporting the move corresponding to the child node population

SDST is illustrated in Fig. 22.7 on the studied game-tree. As for the previously discussed algorithm, a majority of agents in the root node population first points toward the 'move right' (the best move in a purely standard Monte-Carlo sense) before reorienting toward the 'move left' (the best move in the minimax sense). However, the distribution of the agents in the entire metapopulation is now dynamically regulated: most of the agents diffuse in the right part of the game-tree in the first four iterations, and then diffuse back to the left part of the tree in the following iterations. Also, only the regions of interest are visited: for example the entire region after Max's right move at the first ply and Min's right move at the second ply is ignored because the entire subtree leads to a win for Max (no agent becomes active in Min's node population to send inactive agents in this area).

Table 22.2 Stochastic diffusion search applied to trees (SDST)

Initialisation During the initialisation phase, all the agents are allocated to the root node population and their hypotheses are selected randomly among the available moves.

Test During the test phase, complete hypotheses are formed. For each agent X in the root node population, an agent Y in the local population pointed by X's hypothesis is selected. Then an agent in the local population pointed by Y's hypothesis is selected, etc, until the local population pointed by the last agent is empty. Once a hypothesis is formulated, a simulation is run and activities of the agents forming the hypothesis are updated.

Diffusion For each local population, the diffusion phase is divided in three subphases:

1. *Backscattering*: the agents that were not contacted to form a hypothesis go back in the parent node population. In order to preserve the hypotheses distribution among the different moves in the parent node population, a backscattered agent chooses its new hypothesis not randomly but by copying the hypothesis of a chosen agent in that population.
2. *Scattering (by active recruitment)*: every active agent X selects another agent Y at random; if Y is inactive, it is sent in the local population pointed by X's hypothesis. Similarly to the backscattering subphase, in order to preserve the hypotheses distribution in the host node population, the scattered agent selects its new hypothesis not randomly but by copying the hypothesis of a chosen agent in that population (if there are no agents at all in the host node population, then the new hypothesis is chosen randomly).
3. *Internal diffusion (by passive recruitment)*: every inactive agent X selects another agent Y at random; if Y is active, X takes Y's hypothesis.

Under normal conditions, an equilibrium between the scattering and backscattering forces eventually appears, leading to a statistically stable metapopulation. A very interesting property of SDST is that this equilibrium depends on the number of agents used. Asymptotically if enough agents are used, the equilibrium is equivalent to minimax. This is the case of the simulation presented in Fig. 22.7: at iteration 12 the metapopulation stabilises in the left part of the game-tree.

22.5 Conclusion

In the first half of this paper, we demonstrated that a simple application of Monte-Carlo methods to Stochastic Diffusion Search could enable a population of agents (very simple ant-like creatures; our eponymous Aunt Hillary) to play a *strategically informed* game of Hex, although it was also demonstrated that such a system is incapable of *tactically informed* play.

To improve tactical play, it is necessary to facilitate a form of forward planning; the latter half of the paper describes how this can be achieved using (a) a metapopulation of simple ant-like agents to represent a minimax game tree and a novel swarm intelligence heuristic to explore this representation and identify good tactical moves. This heuristic we term Stochastic Diffusion Search applied to Trees (SDST).

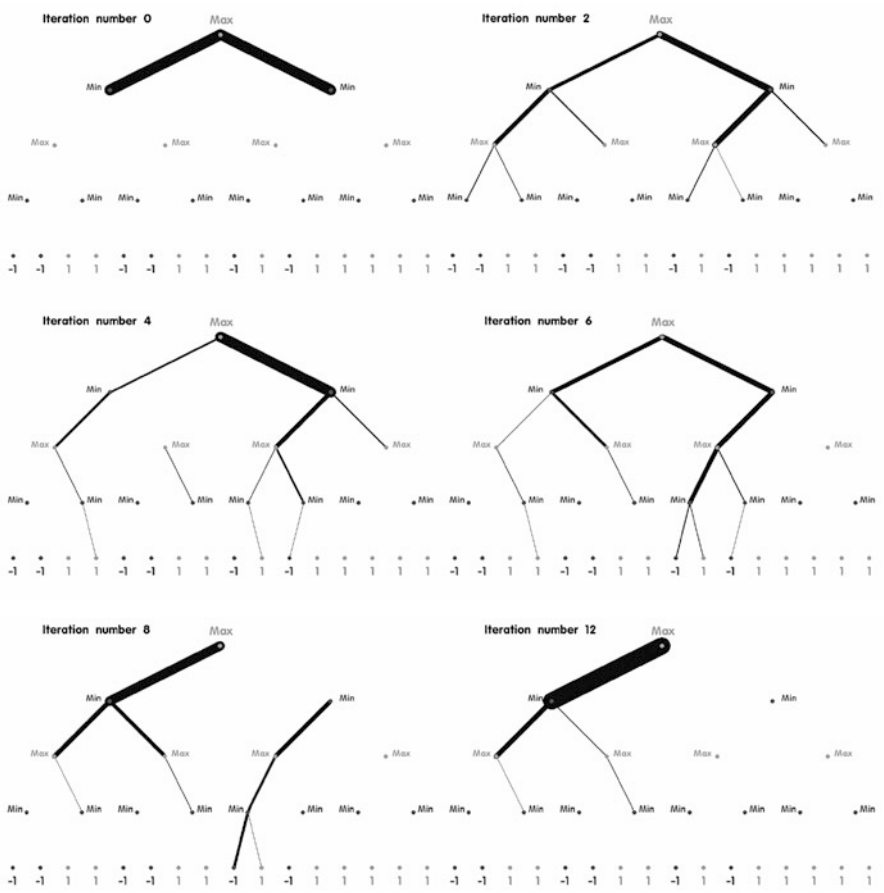


Fig. 22.7 Illustration of SDST: evolution of the distribution of the agents in the entire game-tree (iterations 0, 2, 4, 6, 8 and 12 shown, total number of agents = 100). Each branch has an area proportional to the number of agents in the parent node population supporting the move corresponding to the child node population

SDST is very similar to classical Monte-Carlo Tree Search (MCTS) algorithms in its functioning, but conceptually is radically different. While classical MCTS requires a central processing unit executing the algorithm in a sequential way (with a permanent and complete access to the data), the problem solving ability in SDST emerges from the collaboration of a swarm of homogeneous *ant-like* agents with limited computational capacities.

In addition to developing SDST, our research introduces *meta-level processing* to the Swarm Intelligence paradigm as SDST relies on emergence – both at the level of the agents forming local populations and at the level of the local populations forming a dynamically moving metapopulation. Individual agents are themselves unable to compare the different moves available to them, but their *interaction* leads to the exploitation of the most promising branches at each node of the game-tree.

Similarly, local populations have a weak level of play when taken independently (branches are chosen without tactical sense), but their *interaction* makes a high level of play emerge as SDST is asymptotically equivalent to Minimax. Interestingly, the concept of metapopulation (a population of populations) has been used in biology since 1969 to refer to the dynamical coupling that appears between different populations of social insects (Levins 1969).

Thirdly, the work presented herein takes on its full meaning only if one recognises that it potentially offers interesting insights relating to cognition. In fact, SDS has already been proposed as a model for neural activity: the one-to-one communication makes it a plausible candidate, and there exists a connectionist spiking neuron version of SDS called NESTER (for NEural STochastic nEtwoRk) (Nasuto et al. 2009). Also in SDS, contrary to most of the other swarm intelligence heuristics,[10] semantic interpretation (meaning) is embedded in the entire population of the swarm instead of being simply supported by individual agents.[11] In the neural model NESTER, this property leads to the synchronisation of the firing of neurons at convergence; "*hence in this model oscillatory behaviour may be a result of, rather than a cause of, the binding of features belonging to the same object*" (Nasuto et al. 2009). Furthermore, in addition to offering a novel theoretical solution to the binding problem (Nasuto and Bishop 1998), this ability to efficiently and dynamically allocate cognitive resources in a cognitive search task has been proposed as a model for neural attention (De Meyer et al. 2000).

Finally, in their survey of Monte Carlo Tress Search (Browne et al. 2012), Browne et al. concluded that:

> Over the next five to ten years, MCTS is likely to become more widely used for all kinds of challenging AI problems. We expect it to be extensively hybridised with other search and optimisation algorithms and become a tool of choice for many researchers. In addition to providing more robust and scalable algorithms, this will provide further insights into the nature of search and optimisation in difficult domains, and into how intelligent behaviour can arise from simple statistical processes.

Although it was not conceived for practical AI purposes, we believe that SDST pertains to the type of hybridised algorithm Browne et al. had in mind. In particular, by integrating MCTS with the swarm intelligence paradigm of Stochastic Diffusion Search, we believe that SDST indeed manages to "provide further insights (. . .) into how intelligent behaviour can arise from simple statistical processes."

Acknowledgements The central argument presented herein was developed under the aegis of Templeton project 21853, *Cognition as Communication and Interaction*. The initial development of SDST was extracted from the unpublished MSC Dissertation from Tanay (2012) and from Tanay et al. (2013). This work was originally presented by Bishop at the PT-AI conference St. Antony's College, Oxford, 22nd-23rd September, 2013.

[10] Ant Colony Optimisation also shares this property

[11] This property is due to the partial evaluation of solutions: in the case of string matching for example, as discussed by Nasuto (1999), the position of the solution after convergence is indicated by the formation of a cluster of agents, possibly dynamically fluctuating; in the case of a partial match, agents will keep exploring the text while the cluster will globally stay on the best match.

References

Abramson, B. (1990). Expected-outcome: A general model of static evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(2), 182–193.

Aleksander, I., & Stonham, T. (1979). Guide to pattern recognition using random-access memories. *IEE Journal on Computers and Digital Techniques, 2*(1), 29–40.

Beattie, P., & Bishop, J. (1998). Self-localisation in the 'SENARIO' autonomous wheelchair. *Journal of Intelligent & Robotic Systems, 22*(3), 255–267.

Bishop, J. (1989). Stochastic searching networks. In *First IEE International Conference on Artificial Neural Networks, 1989 (Conf. Publ. No. 313)* (pp. 329–331). IET.

Bishop, J. (1992). The stochastic search network. In R. Linggard, D. Myers, & C. Nightingale (Eds.), *Neural networks for images, speech, and natural language* (pp. 370–387). London/New York: Chapman & Hall.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (2000). Inspiration for optimization from social insect behaviour. *Nature, 406*, 3942.

Browne, C., Powley, E., Whitehouse, D., Lucas, S., Cowling, P., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., & Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games, 4*(1), 1–43.

Chaslot, G., Bakkes, S., Szita, I., & Spronck, P. (2008). Monte-carlo tree search: A new framework for game ai. In *Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference*, Palo Alto (pp. 216–217).

De Meyer, K. (2003). *Foundations of stochastic diffusion search*. Ph.D. thesis, University of Reading.

De Meyer, K., Bishop, J., & Nasuto, S. (2000). Attention through self-synchronisation in the spiking neuron stochastic diffusion network. *Consciousness and Cognition, 9*(2), 81–81.

De Meyer, K., Nasuto, S., & Bishop, J. (2006) Stochastic diffusion optimisation: The application of partial function evaluation and stochastic recruitment in swarm intelligence optimisation. In A. Abraham, C. Grosam, & V. Ramos (Eds.), *Swarm intelligence and data mining* (Vol. 2). Berlin/New York: Springer.

Dorigo, M (1992). *Optimization, learning and natural algorithms*. Ph.D. thesis, Milano: Politecnico di Italy.

Dorigo, M., Maniezzo, V., Colorni, A., Dorigo, M., Dorigo, M., Maniezzo, V., Maniezzo, V., Colorni, A., & Colorni, A. (1991). Positive feedback as a search strategy. Technical report (Technical Report No. 91-016), Politecnico di Milano.

Gale, D. (1979). The game of hex and the Brouwer fixed-point theorem. *The American Mathematical Monthly, 86*(10), 818–827.

Goodman, L. J., & Fisher, R. C. (1979). *The behaviour and physiology of bees*. Oxon: CAB International.

Grech-Cini, H., & McKee, G. (1993). Locating the mouth region in images of human faces. In *Sensor fusion VI* (SPIE-the international society for optical engineering, Vol. 2059). Bellingham: Society of Photo-optical Instrumentation Engineers.

Hart, S. (1992). Games in extensive and strategic forms. *Handbook of Game Theory with Economic Applications, 1*, 19–40.

Hofstadter, D. (1979). *Godel, escher, bach: An eternal golden braid*. New York: Basic Books.

Holldobler, B., & Wilson, E. O. (1990) *The ants*. Cambridge: Springer.

Kennedy J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks (IV)* (pp. 1942–1948).

Kennedy, J. F., Eberhart, R. C., & Shi, Y. (2001). *Swarm intelligence*. San Francisco/London: Morgan Kaufmann.

Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. In *Machine Learning: ECML 2006* (pp. 282–293).

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the ESA, 15*(3), 237–240.

McDermott, D. (2007) Xartificial intelligence and consciousness. In M. Moscovitch, P. D. Zelazo,& E. Thompson (Eds.), *The Cambridge handbook of consciousness*. Cambridge/New York: Cambridge University Press.

Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association, 44*(247), 335–341. doi:10.1080/01621459.1949.10483310. http://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310. PMID: 18139350

Moglich, M., Maschwitz, U., & Holldobler, B. (1974). Tandem calling: A new kind of signal in ant communication. *Science, 186*(4168), 1046–1047.

Nasuto, S. (1999). *Resource allocation analysis of the stochastic diffusion search*. Ph.D. thesis, University of Reading.

Nasuto, S., & Bishop, J. (1998). Neural stochastic diffusion search network-a theoretical solution to the binding problem. In *Proceedings of ASSC2*, Bremen (Vol. 19).

Nasuto, S., & Bishop, M. (1999). Convergence analysis of stochastic diffusion search. *Parallel Algorithms and Applications, 14*(2), 89–107.

Nasuto, S., Bishop, J., & Lauria, S. (1998). Time complexity analysis of the stochastic diffusion search. *Neural Computation, 98*.

Nasuto, S., Bishop, J., & De Meyer, K. (2009). Communicating neurons: A connectionist spiking neuron implementation of stochastic diffusion search. *Neurocomputing, 72*(4), 704–712.

Seeley, T. D. (1995). *The wisdom of the Hive*. Cambridge: Harvard University Press.

Tanay, T. (2012). *Game-tree exploration using stochastic diffusion search*. Technical report, goldsmiths, University of London.

Tanay, T., Bishop, J., Nasuto, S., Roesch E. B., & Spencer, M. (2013). Stochastic diffusion search applied to trees: A swarm intelligence heuristic performing monte-carlo tree search. In *Proceedings of the AISB 2013: Computing and Philosophy Symposium,'What is Computation?'*, Exeter.

Whitaker, R., & Hurley, S. (2002). An agent based approach to site selection for wireless networks. In *Proceedings of the 2002 ACM Symposium on Applied Computing*, Madrid (pp. 574–577). ACM.

Chapter 23
Computer Models of Constitutive Social Practice

Richard Prideaux Evans

Abstract This paper describes a computer implementation of the Game of Giving and Asking for Reasons, as described in *Making It Explicit*. First, I rehearse the distinction between regulative and constitutive views of social practice. It is noteworthy that much multi-agent AI research has been based on the regulative view, despite the philosophical attractions of the constitutive view. Then I distinguish between two sub-types of constitutive interpretation, divided by whether or not intentionality itself is viewed as a bundle of capacities that is constituted by participation in practices. Then I describe a detailed model of the Game of Giving and Asking for Reasons, as a first step in the project of showing how intentionality itself can be realised in a set of practices. I describe the technical points at which I was forced to deviate from Brandom's original description.

Keywords Conversation games • Turing test • Social practice • Multi-agent simulation • Modeling • Inference • Robert Brandom

23.1 The Constitutive View of Social Practices

I shall start by outlining the regulative view of social practices, before describing an alternative: the constitutive view. Then I will distinguish between two sub-types of constitutive view, before describing a computer model of the second sort of constitutive view.

R.P. Evans (✉)
Imperial College, London
e-mail: RichardPrideauxEvans@gmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_23

23.1.1 The Regulative View of Social Practice

Research in multi-agent systems typically[1] assumes a *regulative* model of social practice. In this model, the agent starts with a given set of goals, and a given set of actions for achieving these goals. Social practices are then introduced afterwards as a way of achieving coordination when there are multiple agents whose activity can come into conflict. In this regulative model, a social practice provides a *restriction* on the set of (antecedently given) actions so as to satisfy the (antecedently given) set of goals. For example, in a world containing cars but no driving regulations, agents are free to drive on either side of the road. To prevent collisions, we introduce driving regulations, insisting that everyone drives on the left hand side of the road. We accept this limitation on our freedom because it helps us satisfy our goal of survival.

23.1.2 The Constitutive View of Social Practice

The alternative *constitutive* view of social practices denies that actions or goals can always be specified in advance, prior to and independently of the practices the agent is participating in. The constitutive view claims that there are some actions that are *only available* because one is participating in a social practice with a certain structure. For example:

- You can swing a peculiarly shaped piece of wood without participating in any particular practice – but this action will only constitute a *strike* if you are participating in a game of baseball
- I can raise my hand whenever I like, but this only counts as *voting for the motion* within the institution of voting
- I am free to say "I do" at any moment, but these sounds only constitute *getting married* within a wedding ceremony

Now there are some actions I cannot perform alone because of some *contingent* fact about myself. So, for example, I cannot move the sofa on my own. But perhaps, if I was stronger or more manly, I could. The constitutive claim is much stronger than that: the claim is that there are certain actions which *necessarily* cannot be done unless they are achieved within a particular practice.

Another point of clarification: the new action that the practice enables is not a new type of *physical* action. It isn't like the case of Spiderman, who was bitten by a radioactive spider and suddenly had new physical capacities (e.g. the ability to climb walls). Rather, the new actions that are made available by practices

[1] See Moses and Tenenholtz (1992) or Shoham (2008). For an explicit acknowledgement of the importance of constitutive conditionals in multi-agent simulations, and the beginning of an analysis, see Jones and Sergot (1996).

are *re-interpretations* of previously-achievable actions. Using Searle's counts-as formulation (Searle 1992):

x counts as y in context c

The social practice provides a context c in which the already-achievable action x is now also the performance of y. So, to take a very well-worn example, moving the piece of wood from one square to another *counts as* moving the knight to king's-bishop 3 in the context of the game of chess.

23.1.3 Two Types of Constitutive View

The initial formulations of the constitutive view[2] imagine an agent who is *already* intentional *before* he participates in practices. This *amplificative* interpretation of the constitutive view pictures an agent who is already able to represent, reason and plan. The practices provide him with a way of *expanding* his capacities and goals.

But this work is based on a stronger, more controversial understanding of constitutive social practice. According to the *foundational view*, the agent's abilities to represent, reason and plan are themselves constituted by the practices he is participating in. *Intentionality itself* is one of the (clusters of) capacities that social practices enable.

Consider Searle's distinction in Searle (1969) between *intrinsic* and *derived* intentionality. According to him, people have intrinsic intentionality, while pieces of paper (and computers) have merely derived intentionality: the fact that the writing on the piece of paper means something is only true because there are agents who (intrinsically) mean something by that same sentence when they produce/interpret it. The stronger, more controversial, understanding of constitutive practice claims that the intentionality of *people* is also derivative. The only thing that has intrinsic intentionality is the *practice*.

More specifically, when an agent means something by a sentence, that is only because he is participating in a practice in which that sentence means what it does. This is an example of a *mediating relation*.

23.1.3.1 Mediating Relations

Some relations can be best understood by splitting them in two, and inserting an intermediate entity between the relata. In these cases, the explanation is of the form:

$$\forall x, y.\ A(x, y) \text{ because } \exists z.\ B(x, z) \wedge C(z, y)$$

[2]See Rawls (1955) and Searle (1969). Even in these early formulations, there were important differences: Rawls distinguishes between statistical regularities and constitutive rules, while Searle distinguishes between constraining(regulative) rules and constitutive rules.

For example:

- x is the aunt of y because there is a person z who is the sibling of x and the parent of y
- x is a logical (proof-theoretic) consequence of y iff there is a proof z such that z starts with y and ends with x
- x is married to y because there has been a wedding ceremony z in which x was groom in z and y was bride in z

One of the central claims in *Making It Explicit* (Brandom 1998) is that intentional states should be explained in terms of participation in practices:

> Expressions come to mean what they mean by being used as they are in practice, and intentional states and attributes have the contents they do in virtue of the role they play in the behavioral economy of those to whom they are attributed.

This is another example of a mediating explanation. Having an intentional state is a relation between an agent and an intentional state which is best understood via a mediating entity – a social practice:

- x has intentional state y because there is a practice z such that x participates in z and z institutes y

23.1.4 Summary

So far, I have outlined three levels of increasing commitment to social practices:

1. Social practices as *restrictions* on (antecedently given) actions to satisfy (antecedently given) goals
2. Social practices as *amplificative*: ways of providing extra actions and goals to agents who are already intentional (capable of representing, reasoning, and planning)
3. Social practices as *foundational*: providing the structure needed for all intentional activity

To place some names on these positions:

1. Much work in multi-agent AI research (e.g. Moses and Tenenholtz 1992; Shoham 2008) work from the first assumption
2. Searle (1992) and Rawls (1955) work from the second assumption
3. Brandom (1998) explicitly argues for the third position

The computer model described below starts from the third, strongest, most controversial assumption. If we want to evaluate the claim that activity in general is made possible through social practice, then we should start with the hardest case first. The hardest case is intentionality. If we can make it plausible that intentionality itself can be instituted through social practice, then the other cases will be relatively straightforward.

23.2 Modelling the GOGAR

23.2.1 Introduction

The GOGAR (**G**ame **O**f **G**iving and **A**sking for **R**easons) is a social practice which enables participants to make assertions by producing sounds. If a parrot says "Nice weather we're having", she has not *asserted* that the weather is nice – she has merely produced some noises. But if an agent utters these sounds as a participant in the GOGAR, it counts as *asserting* that the weather is nice. GOGAR is the practice which turns sounds into assertions.

The GOGAR keeps track of who has said what (the commitments). It divides the commitments into two groups: those that are "entitled",[3] and those that are not.

Claims are entitled by default, but a claim can lose its entitlement if it is challenged. A claim can be challenged by a non-propositional speech-act (the incredulous raising of an eyebrow), or by a propositional speech-act: by asserting a proposition which is incompatible with it.

When a claim is challenged, it loses its entitlement. But it can get its entitlement reinstated if it is justified by other assertions. These justifications are themselves just other assertions, which can themselves be challenged by other incompatible assertions – in which case, the justifications will themselves need further justifications to reinstate their entitlement – and these further justifications can themselves be challenged, and so on.

The rest of this section will describe a computer model of the GOGAR. The description of the GOGAR in Brandom (1998) is remarkably precise, making it relatively straightforward to implement what he described. But we will see that we will have to depart from Brandom at one crucial point.

23.2.2 Basic Definitions

Given a background set $\mathcal{X}$ of agents and $\mathcal{S}$ of sentences, then a debate-state $\mathcal{D}_x$ according to a particular agent $x \in X$ is a tuple $(\mathcal{A}, \mathcal{C}, \mathcal{E}, \mathcal{I})$, consisting of:

1. A set $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{S}$ of assertions. Each assertion is a pair consisting of the the person who asserted it and the sentence asserted.
2. A set $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{P}(\mathcal{S})$ of commitment-preserving inferences in horn-clause form. Each horn clause $A \leftarrow C_1 \wedge \ldots \wedge C_n$ is represented by the pair $(A, \{C_1, \ldots, C_n\})$.

[3] Brandom uses "entitlement" in a more-or-less epistemic sense. A sentence is entitled if the speaker is *justified* in asserting it. But in the discussion that follows, I provide reasons to lose these individual epistemic associations. So think of "entitlement" simply as a term of art, describing a property that asserted sentences *should* have (and if they don't, the speaker should justify or retract the claim). We are interpreting entitlement as an abstract social status with normative consequences – rather like the property of being-in-check in chess.

3. A set $\mathcal{E} \subseteq \mathcal{S} \times \mathcal{P}(\mathcal{S})$ of entitlement-preserving inferences in horn-clause form.
4. A set $\mathcal{I} \subseteq \mathcal{P}(\mathcal{P}(\mathcal{S})$ of sets of incompatibility sets.

Different agents will have different interpretations of the same debate. The set $\mathcal{A}$ of assertions will vary between different agents: one might not have heard an assertion if he misheard, was out of earshot, or was not paying attention. Different agents may have different understandings of the inferential relations: one agent may think that p is a commitment-preserving consequence of q, while another may not. There can be similar disagreements about entitlement-preserving relations and incompatibility relations: one agent may think that, for example, free will and determinism are incompatible propositions, while another may think they are compatible.

At the heart of the GOGAR is a function which computes, given a particular understanding of the debate $\mathcal{D}_x = (\mathcal{A}, \mathcal{C}, \mathcal{E}, \mathcal{I})$ according to a particular agent x, the subset of assertions which x thinks are entitled:

$$\text{Entitled}_x(\mathcal{A}, \mathcal{C}, \mathcal{E}, \mathcal{I}) \subseteq \mathcal{A}$$

All claims are entitled by default. A claim loses its entitlement if it is challenged. The entitled claims are claims that either remain unchallenged – or have been challenged, but have also been successfully justified. To say that a claim is not entitled is to make a claim with normative consequences: the claims that are not entitled are the ones that the asserter *should* justify (or retract).

23.2.3 Incompatibility, Commitment and Entitlement

In *Making It Explicit* (p. 194), Brandom says

> Two claims are incompatible if commitment to one precludes entitlement to the other

But incompatibility is a relation between sentences, while commitment is a relation between a speaker and a sentence. If p and q range over sentences, and x and y range over agents, there are two possible interpretations of p and q being incompatible:

1. If x is committed to p, then y cannot be entitled to q
2. If x is committed to p, then x cannot be entitled to q

In the first interpretation, one speaker's assertion can affect the entitlement of another speaker's assertion. In the second interpretation, one speaker is making both assertions.

The first interpretation is too strong. It means that if x asserts p, I can prevent x ever being entitled to p just by asserting an incompatible proposition q. This interpretation is clearly is not what Brandom means.

The second interpretation is the one that Brandom uses.[4] But it is, I submit, too weak. Suppose x asserts p and y asserts q, where p and q are incompatible. According to this second interpretation, both x's claim that p and y's claim that q can *both* retain their entitlement, even though they are incompatible. This means that the assertions of one speaker can have no effect whatsoever on the entitlement of another speaker's assertions! They are closed off from affecting each other. But one person's claim can only count as a *challenge* to another person's claim if the former can affect the entitlement of the latter.

So, in one way, our second interpretation is much too weak. But in another way it is too strong. Suppose x asserts p and q (where again p and q are incompatible). Suppose further that x provides some very compelling arguments for p, but none for q. Now he is committed to q, so if commitment to q precludes entitlement to p, then he can never get entitlement reinstated for p, *no matter how compelling* the justifications he provides for p. The only way, according to this interpretation, that he can get entitlement for p is if he *retracts* q.

The second interpretation is the one that MacFarlane implemented in his own implementation of GOGAR (MacFarlane 2006). But his implementation has exactly the issue outlined above, that different speakers' claims cannot affect each others' entitlements:

```
Welcome to the game of giving and asking for reasons,
a simulation of the linguistic scorekeeping dynamics
described in chapter 3 of Robert Brandom's book
Making It Explicit (Harvard University Press, 1994).

(c) 2006 John MacFarlane

For a list of sample commands, type help

GOGAR> Bob asserts A is red
GOGAR> Ann asserts A is blue

Bob's score on Ann
Commitments:  {A is colored, A is blue}
Entitlements: {A is colored, A is blue}
Incompatibles:

Bob's score on Bob
Commitments:  {A is red, A is colored}
Entitlements: {A is red, A is colored}
```

[4] See Brandom (2008, p. 120): Incompatibility of p and q is defined as "If S is committed to p, then S is not entitled to q". This is also the interpretation John MacFarlane uses in MacFarlane (2006).

```
Incompatibles:

GOGAR> Bob asserts A is blue
Bob's score on Ann
Commitments:  {A is colored, A is blue}
Entitlements: {A is colored, A is blue}
Incompatibles:

Bob's score on Bob
Commitments:  {A is red, A is colored, A is blue}
Entitlements: {}
Incompatibles: {A is red, A is blue}
```

First Bob asserts "A is red". Then Ann asserts "A is blue".[5] At this point, despite the fact that they have challenged each other with directly incompatible assertions, neither of the two claims has lost entitlement. Finally, Bob asserts "A is blue". It is only when a speaker contradicts *himself* that he loses entitlement. This simple example shows the problems with this second interpretations: direct challenges go entirely unnoticed.

As neither of the above interpretations are satisfactory, we are compelled to look elsewhere. The simplest alternative is to explain incompatibility between p and q as:

- If x is entitled to p, then y cannot be entitled to q

Two claims are incompatible if entitlement to one precludes entitlement to the other (no matter who said them).

The difference between this interpretation and Brandom's is that he sees entitlement as a collective version of *justification*: just as two people can believe incompatible claims, and both be justified in their beliefs, just so two people can be entitled to their incompatible assertions. In the proposed alternative interpretation, by contrast, entitlement is a collective version of *knowledge*. Imagine a social practice in which the monadic predicate "It is known that p" is prior to the dyadic "x knows that p". In this alternative interpretation, entitlement to p represents "It is known that p". Since two incompatible propositions cannot both be known, two incompatible propositions cannot both be entitled.

In the rest of this section, I describe a computer implementation of GOGAR which assumes this alternative interpretation of the relation between incompatibility and entitlement.

[5]The implicit assumption is that A is monochromatic.

23.2.4 Computing Entitlement

To ease exposition, we make two simplifying assumptions:

1. We ignore who said what, treating an assertion as a simple sentence (ignoring who uttered it)
2. We ignore the commitment-preserving inferences, and just assume the set of assertions is closed under the commitment-preserving relation

So now a debate $\mathcal{D}_x$ is just a triple $(\mathcal{A}, \mathcal{E}, \mathcal{I})$, consisting of:

1. A set $\mathcal{A} \subseteq \mathcal{S}$ of assertions
2. A set $\mathcal{E} \subseteq \mathcal{S} \times \mathcal{P}(\mathcal{S})$ of entitlement-preserving inferences in horn-clause form
3. A set $\mathcal{I} \subseteq \mathcal{P}(\mathcal{P}(\mathcal{S})$ of sets of incompatibility sets

Example 1. For example

$$\mathcal{A} = \{p, \neg p, q\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

Figure 23.1 shows the state of the debate in Example 1. Arrows indicate entitlement-preserving rules and dotted boxes represent incompatibility sets. Entitled claims are drawn in circles.

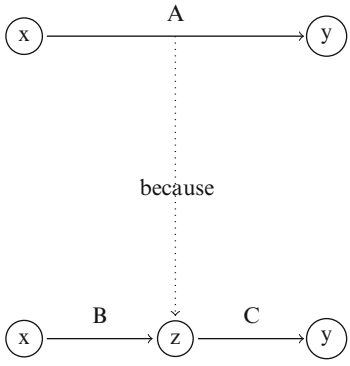


Fig. 23.1 Mediating explanations

Fig. 23.2 Example 1

Fig. 23.3 Example 2

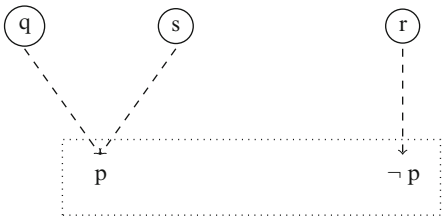


Fig. 23.4 Example 3

Example 2.

$$\mathcal{A} = \{p, \neg p, q, r\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, \neg p \leftarrow \{r\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

Example 3.

$$\mathcal{A} = \{p, \neg p, q, r, s\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, p \leftarrow \{s\}, \neg p \leftarrow \{r\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

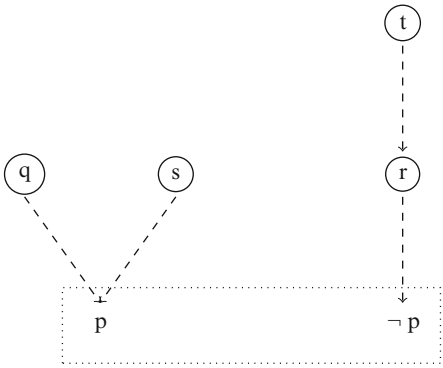
Example 4.

$$\mathcal{A} = \{p, \neg p, q, r, s, t\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, p \leftarrow \{s\}, \neg p \leftarrow \{r\}, r \leftarrow \{t\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

Computing the entitled claims involves, at the procedural level:

1. Start by assuming all the claims are entitled
2. Remove entitlement from all incompatible subsets
3. Compute the immediate consequences of the currently entitled claims
4. Remove entitlement from all incompatible subsets
5. Repeat steps 3 and 4 until no more propositions are added to the set of entitled claims

Fig. 23.5 Example 4



Redescribing this procedure at the functional level, let

$$\text{Inc}(s) = \{x \in s \mid \exists y_1, \ldots, y_n \in s \text{ such that } \{x, y_1, \ldots, y_n\} \in \mathcal{I}\}$$

Let

$$\phi(s) = s - \text{Inc}(s)$$

Define $Cn_1(s)$ as the set of immediate consequences of s according to the horn-clauses in $\mathcal{E}$:

$$Cn_1(s) = s \cup \{p \in s \mid (p \leftarrow q) \in \mathcal{E} \wedge q \subseteq s\}$$

Now define a function $N : \mathcal{P}(T) \rightarrow \mathcal{P}(T)$:

$$N = \phi \cdot Cn_1$$

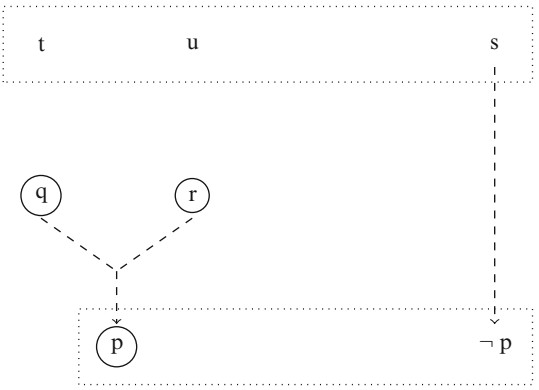
Note that ϕ and N are not monotonic. Now define a sequence of entitlement sets $E_0, E_1, \ldots$. where:

$$E_0 = \phi(\mathcal{A})$$
$$E_{n+1} = N(E_n)$$

Now, although N is not monotonic, we have $E_n \subseteq E_{n+1}$ for all n. So, if $\mathcal{S}$ and $\mathcal{E}$ are finite, this sequence converges, and we define

$$\text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) = \bigcup_{i \geq 0} E_i$$

Fig. 23.6 Example 5



Example 5. Given:

$$\mathcal{A} = \{p, \neg p, q, r, s, t, u\}$$
$$\mathcal{E} = \{p \leftarrow \{q, r\}, \neg p \leftarrow \{s\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}, \{t, u, s\}\}$$

The computation of Entitled involves:

$$\begin{aligned} E_0 &= \{q, r\} \\ E_1 &= \{q, r, p\} \\ E_2 &= \{q, r, p\} \\ &\ldots \\ \text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) &= \{q, r, p\} \end{aligned}$$

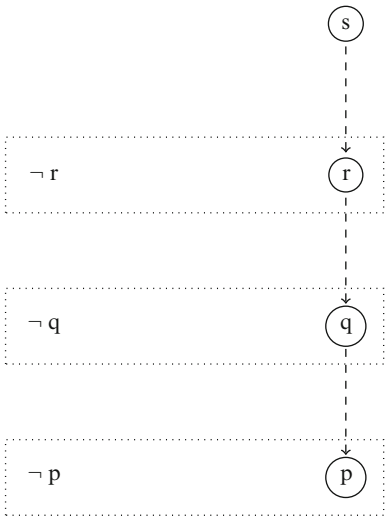
Example 6. A slightly more complex example, consider:

$$\mathcal{A} = \{p, \neg p, q, \neg q, r, \neg r, s\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, q \leftarrow \{r\}, r \leftarrow \{s\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}, \{q, \neg q\}, \{r, \neg r\}\}$$

The computation of Entitled involves:

$$\begin{aligned} E_0 &= \{s\} \\ E_1 &= \{s, r\} \\ E_2 &= \{s, r, q\} \end{aligned}$$

Fig. 23.7 Example 6



$$
\begin{aligned}
E_3 &= \{s, r, q, p\} \\
E_4 &= \{s, r, q, p\} \\
&\ldots \\
\text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) &= \{s, r, q, \neg p\}
\end{aligned}
$$

23.2.4.1 Entitlement-Preserving Inferences Do Not Always Preserve Entitlement in the Presence of Conflicting Claims

One noteworthy aspect of this way of computing entitlement is that it allows the possibility that the following are simultaneously true:

- There is an entitlement-preserving inference from q to p: $p \leftarrow \{q\} \in \mathcal{E}$
- q is entitled: $q \in \text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I})$
- p is not entitled: $p \notin \text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I})$

For example, consider the following case:

Example 7.

$$
\begin{aligned}
\mathcal{A} &= \{p, \neg p, q, r\} \\
\mathcal{E} &= \{p \leftarrow \{q\}, \neg p \leftarrow \{r\}\}\} \\
\mathcal{I} &= \{\{p, \neg p\}\}
\end{aligned}
$$

Fig. 23.8 Example 3

Here, the entitled claims are:

$$\text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) = \{q, r\}$$

Now, it might seem odd (or worse, plain wrong) to allow an entitlement-preserving inference that does not automatically transfer entitlement from premise to conclusion (and contrapositively, transfer lack of entitlement from conclusion to premise). But this is to misunderstand the nature of an entitlement-preserving inference: it does preserve entitlement by default (see Cn_1) – but only *in the absence of countervailing factors*. This is a non-monotonic logic: other commitments may preclude entitlement from the conclusion even if the premise is entitled.[6]

23.2.4.2 Entitlement as an Abstract Normative Status

Entitlement, as defined here, is an abstract property which is used to determine which statements require justification. The way entitlement is defined abstracts away from many of the historical details of how the claims were uttered:

- Making the same statement *more than once* has no effect whatsoever on the set of entitled claims
- The *order* in which claims are made has no effect whatsoever on the set of entitled claims
- *Who* is making which claims has no effect whatsoever on the set of entitled claims. If you claim p and $\neg p$, then entitlement-wise this is no different from if you claim p and I claim $\neg p$: in both cases, both claims lack entitlement

[6] Brandom makes this point in Brandom (2008): "As will appear, entitlement-preserving inferences are always defeasible; the entitlement one acquires thereby is only *prima facie*."

23.2.4.3 Computing the Extra Assertions Needed to Make a Claim Entitled

Suppose, in a debate, a certain claim lacks entitlement and we want to reinstate its entitlement. Typically there will be various different sets of claims one could add which would restore its entitlement. These sets can be calculated as follows. Given a debate $\mathcal{D} = (\mathcal{A}, \mathcal{E}, \mathcal{I})$ involving a total set $\mathcal{S}$ of propositions, the possible justifications $\mathcal{J}$ of a claim p are defined as:

$$\mathcal{J}(p) = \{c' \in \mathcal{P}(\mathcal{S} - \mathcal{A}) \mid p \in \text{Entitled}(c' \cup \mathcal{A}, \mathcal{E}, \mathcal{I}) \wedge \neg\, \exists c'' \in \mathcal{P}(\mathcal{S} - \mathcal{A})\ c'' \subset c'\}$$

In other words, the possible justifications of p are the minimal subsets c' of $\mathcal{S}$ that can be added to $\mathcal{A}$ such that p is entitled in $(c' \cup \mathcal{A}, \mathcal{E}, \mathcal{I})$.

Example 8. Suppose we have

$$\mathcal{A} = \{p, \neg p, q, \neg q, r, \neg r, s, t, u\}$$
$$\mathcal{E} = \{p \leftarrow \{q, r\}, \neg p \leftarrow \{s\}, \neg p \leftarrow \{t, u\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}, \{q, \neg q\}, \{r, \neg r\}\}$$

Then the various justifications of $\neg p$ are:

$$\mathcal{J}(\neg p) = \{\{\neg q, s\}, \{\neg r, s\}, \{\neg q, t, u\}, \{\neg r, t, u\}\}$$

It is a consequence of the definition of entitlement that a debate can never get into a hopeless position. Any claim in any debate is *redeemable* in the sense that for all propositions p that are claimed in debate $(\mathcal{A}, \mathcal{E}, \mathcal{I})$ involving propositions $\mathcal{P}$, there are extensions $(\mathcal{A}' \supseteq \mathcal{A}, \mathcal{E}' \supseteq \mathcal{E}, \mathcal{I}' \supseteq \mathcal{I})$ in $\mathcal{P}' \supseteq \mathcal{P}$ such that p is entitled in $(\mathcal{A}', \mathcal{E}', \mathcal{I}')$.

23.3 Modelling Pragmatic Factors

23.3.1 The Burden of Proof

One norm lies at the heart of the GOGAR:

> If x is committed to p, and p lacks entitlement, then x should either justify or retract p.

In situations where two agents have made incompatible assertions, neither of which is justified, they *both* have to justify (or retract) their claims. This model contrasts with the traditional concept of the "burden of proof". According to this concept, there is at most one person who has the "burden of proof" at any moment. It may be sometimes difficult to assess who has it, and different people may disagree, but nevertheless there is at most one person who can have it.

The model proposed above is more equanimous. There are often cases (see Examples 2, 3 and 4) where multiple agents have to justify their claims. There is no single agent who the burden of proof falls on. In these cases, any attempt to single out an individual as particularly responsible is an attempt to impose a *power relation* on an essentially symmetrical situation.

23.3.2 Norms and Power Relations

At the heart of the GOGAR lies the norm:

> If *x* is committed to *p*, and *p* lacks entitlement, then *x* should either justify or retract *p*.

But norms do not work by magic. They need to be articulated, monitored and enforced by the activities of individual agents.

When there is a norm that *x* should justify *p*, who should articulate, monitor and enforce that norm? There are two broad ways this can happen:

- Another *y* can insist that *x* justify his claim by playing *high-status* to *x*. (For example, *y* says to *x*: "You really need to justify your claim that *p*").
- *x* himself can enforce the norm by playing low-status to another *y*. (For example, *x* says to *y*: "Oh dear, my claim that *p* has been challenged. I really need to justify it").

The terms *high status* and *low status* come from Keith Johnstone's *Impro* (Johnstone 1987). These are power relations in a short-term local situation.[7] In Johnstone's sense, a pauper can play high-status to a king while he is bossing him around. (Humour often arises from unexpected status games).

The high/low status-game is implemented as a separate social practice, running concurrently with the GOGAR, which monitors the state of the GOGAR and provides appropriate status-related affordances. The status-game can be in one of three states:

- Neutral: no status game is currently being played
- High/Low(*x*,*y*): agent *y* is playing high to agent *x*
- Conflict(*x*, *y*): agent *x* and *y* are in conflict as a result of trying to play high-status to each other

There are different affordances available in the different states. In the Neutral state, whenever there is a claimer *x* who has claimed something which is not entitled, there is an opportunity for

- *x* to play low to anyone
- another *y* to play high to *x*

[7] As opposed to social status (e.g. upper class) which are long-term and (at least in some societies) difficult to change within one's own life-time

In the High/Low(x,y) state, the affordances available include:

- y remind x that x needs to justify his/her claim
- x look worried
- x get annoyed with y (if x does not like being low status)

So, for example, if x and y have made incompatible assertions p and q, and neither p nor q has been justified, then they can both play high-status on each other. Suppose y plays high status to x. Now if x is also a habitual high-status player, he will not enjoy being made to play low-status, and will get annoyed.

23.3.3 Turn-Taking and Conversational Salience

In an earlier implementation of GOGAR, the agents had a good understanding of when their claims had been challenged, and what sort of claims to make to justify their assertions – but they had no understanding of the conversational context in which their claims were embedded. If an agent had one of his claims challenged earlier, but then the conversation had moved on, the agent would go back, relentlessly, to his previously challenged assertion. He had no understanding of whose turn it was to speak, or the current focus (or foci) of the conversation.

To address this, our latest implementation combines a model of giving-and-asking-for-reasons with a model of conversational salience (based on Sacks, Schegloff and Jefferson's seminal paper on conversational turn-taking Sacks et al. 1974). Conversational turn-taking is itself modelled as a social practice, containing norms describing who should speak next, on what topic, and with various mechanisms for repairing conversational blunders. Now, the agents understand both what has been said and *when it is appropriate to respond.*

The turn-taking practice involves two core concepts:

- the selected speaker (if any)
- the selected topic (or topics)

The selected speaker is the agent (if any) who should speak next. Certain types of utterance directly determine a selected speaker. For example:

- x says to y: "Can you please stop standing on my foot?"

Others indirectly determine a selected speaker. If x has previously asserted p and y makes a claim which is incompatible with q, then this is a challenge to x's previous claim and x is the selected speaker. But sometimes, there is no selected next speaker and anybody can speak next. For example:

- x (addressing the group in general): "Has anybody seen my hat?"

Each utterance is about one or more topics. The most recent utterance determines the selected topics.

The turn-taking practice enforces the following rules:

- If there is a selected speaker, then he should speak next. Other people speaking out of turn constitutes an interruption.
- If there is no selected speaker, anyone may speak next.
- The next utterance should involve one of the selected topics. If the selected speaker cannot continue any of the selected topics, he must preface his utterance with a preamble connecting it to a previous topic, or preface it with a conventional way of clearing selected topics ("anyway…"). Failure to respect this rule constitutes an interruption.

23.4 Evaluation, Limitations and Further Work

Austin once wrote (Austin 1956):

> In the history of human inquiry, philosophy has the place of the central sun, seminal and tumultuous: from time to time it throws off some portion of itself to take station as a science, a planet, cool and well regulated, progressing steadily towards a distant final state.

He believed that one day,[8] the heated debates surrounding philosophy of language would cool down, and the hard-won accumulated insights would start to build into a science of language.[9]

The computer model described here started by taking seriously the idea that philosophical insight can serve as direct inspiration for AI architectures. A number of the fundamental architectural features were based on (controversial) philosophical insights. I have already dwelt at length with three such claims:

- certain actions are constituted in certain practices
- certain goals are constituted in certain practices
- intentionality, in particular, is constituted in a particular practice (GOGAR)

I shall briefly describe one other philosophical claim that fundamentally informed the architecture. Typically multi-agent simulations of social practices model the world as a collection of objects.[10] This simulation, by contrast, takes the Tractatus idea seriously that the world is *everything that is the case* (Wittgenstein 1961): the entire simulation state is defined as a set of sentences in a formal language. Choosing a declarative representation of simulation state has been shown to have certain significant practical advantages when building a complex simulation: it is

[8] He believed it would happen during the twenty-first century.

[9] When he said a science of language, he did not just mean a science of formal linguistics (a la Chomsky and formal grammar), but a science of language *use*.

[10] In an object-oriented representation, each object is a cluster of facts, related to other objects via pointers.

significantly easier to visualise, debug, and serialise the simulation state.[11] Time and again, hard-won philosophical insights informed the fundamental architectural decisions.

23.4.1 Limitations and Further Work

So far, I have built simulations of simple debates using the architecture described above. But these are only toy examples and initial explorations. There is a huge amount more to do. I will focus on two aspects in particular:

- supporting language entry and exit moves
- agents making their inferential relations explicit

23.4.2 Supporting Language Entry and Exit Moves

Sellars (1954) coined the term "language entry move" for an inference from a perception to an assertion. He used the term "language exit move" for a transition from an assertion to an action. The computer model of GOGAR implemented so far has neither language-entry nor language-exit moves.

- Language-entry moves. The agents start off with a given set of beliefs. This set does not expand or contract during simulation.
- Language-exit moves. The computer agents do not *act* on their beliefs (apart from asserting or justifying them).

A richer simulation would model the way agents acquire information, how a debate can change someone's mind, and how their beliefs can affect their subsequent actions (as Marx famously insisted on).

23.4.3 Making Inferential Relations Explicit

In the GOGAR, agents typically have different understandings of the inferential relations. They will have different understandings about which claims are incompatible, commitment-preserving, and entitlement-preserving. Their different understandings of the inferential relations leads to different understandings of the

[11] The formal language used to represent the world was not traditional predicate logic, but a modal language designed to model social practices. This modal logic (Eremic Logic) is itself inspired by the Sellars/Brandom thesis that material incompatibility is prior to logical negation. In this logic, you can express directly the fact that "x is blue" and "x is red" are incompatible.

entitlements. For example, suppose *x* believes free-will and determinism are incompatible, while *y* believes they are compatible. Consider the following exchange:

1. *x*: 'We have free-will'
2. *y*: 'All events are entirely determined'
3. *x*: 'Yes, that is also true'

By the third claim, at the point at which *x* agrees with *y*, *x* (the compatibilist) believes both 1 and 2 are entitled. *y* (who is an incompatibilist) thinks both claims have lost entitlement.

In the current implementation, such differences cannot be resolved because they cannot be made explicit. But consider one possible continuation:

1. *y*: 'Huh? You have a contradicted yourself'
2. *x*: 'No I haven't'
3. *y*: 'Yes you have. You claimed that we have free well, and that all events are entirely determined. These two claims are incompatible.'
4. *y*: 'No – they are not incompatible.'
5. *x*: 'Yes they are. It is part of the meaning of an event being freely willed that you could have done otherwise – but if determinism is true, you could not have done otherwise.'

In the continuation, the debate is raised to the meta-level. Instead of disagreeing about first-order issues (free-will versus determinism), they are now disagreeing about whether *x* has contradicted himself, and whether or not two claims are incompatible. There is an incompatibility at the meta-level between the claims "*x* has contradicted himself" and "*x* has not contradicted himself". There is also an incompatibility between "*p* and *q* are compatible" and "*p* and *q* are incompatible". By the 7th claim, *y* has challenged the incompatibility of free-will and determinism. This incompatibility claim (which was previously implicit in *x*) is now challenged.

Recall the first-level norm:

> If *x* is committed to *p*, and *p* lacks entitlement, then *x* should either justify or retract *p*.

There is a related norm at the meta-level:

> Once an inferential relation has been made explicit and lost entitlement, it should no longer be used in inference until its entitlement is restored.

So *x* is stymied until he can provide justification for the incompatibility. It is only when he makes the 8th claim that entitlement is restored.

Meta-level debating involves making the various inferential relations explicit. We have seen an example where an incompatibility relation was made explicit. Here is an example where an entitlement-preserving relation is made explicit and challenged:

1. *x*: *p*
2. *y*: ¬*p*
3. *x*: *q*
4. *y*: "You need to provide a justification for *p*."

5. *x*: "No I don't - *p* is already justified by *q*."
6. *y*: "On the contrary - *q* may be true but it does not support *p*."

Here, they disagree about whether *p* is entitled, and then proceed to disagree about whether *q* is an entitlement-preserving reason for *p*.

When we add this sort of meta-level debating to the GOGAR, agents will be able to change their mind – not only about what they believe – but also about what they *mean*.

References

Austin, J. L. (1956). Ifs and cans. *Proceedings of the British Academy, 42,* 109–132.
Brandom, R. (1998). *Making it explicit*. Cambridge: Harvard University Press.
Brandom, R. (2008). *Between saying and doing*. Oxford/New York: Oxford University Press.
Johnstone, K. (1987). *Impro*. New York: Routledge.
Jones, A., & Sergot, M. (1996). A formal characterisation of institutionalised power. *Logic Journal of the IGPL, 4*(3), 427–443.
MacFarlane, J. (2006). *GOGAR*. http://johnmacfarlane.net/gogar.html
Moses, Y., & Tenenholtz, M. (1992). On computational aspects of artificial social systems. In *Proceedings of 11th DAI Workshop*, Glen Arbor.
Rawls, J. (1955). Two concepts of rules. *The Philosophical Review, 64*(1), 3–32.
Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50,* 696–735.
Searle, J. R. (1969). *Speech acts*. London: Cambridge University Press.
Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge: MIT.
Sellars, W. (1954). Some reflections on language games. *Philosophy of Science, 21*(3), 204–228.
Shoham, Y. (2008). *Multiagent systems*. Cambridge/London: Cambridge University Press.
Wittgenstein, L. (1961). *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul.

Part IV
Embodied Cognition

Chapter 24
Artificial Intelligence: The Point of View of Developmental Robotics

Jean-Christophe Baillie

Abstract We present here the research directions of the newly formed Artificial Intelligence Lab of Aldebaran Robotics. After a short historical review of AI, we introduce the field of developmental robotics, which stresses the importance of understanding the dynamical aspect of intelligence and the early developmental stages from sensorimotor categorization up to higher level socio-cognitive skills. Taking inspiration in particular from developmental psychology, the idea is to model the underlying mechanisms of gradual learning in the context of a progressively more complex interaction with the environment and with other agents. We review the different aspects of this approach that are explored in the lab, with a focus on language acquisition and symbol grounding.

Keywords Aldebaran • Robotics • Developmental robotics • Learning • Symbol grounding

24.1 The A-Labs

Aldebaran Robotics is a French company specialized in humanoid robotics with a prime focus on human-robot interaction. The main product line (Nao and Romeo robots at the moment) are currently sold as programmable platforms but the long term goal of the company is to offer highly interactive and socially capable robots for everyday use, whether in the framework of assistive robotics, entertainment or education.

To advance this goal the company has invested heavily in research related to artificial intelligence and robotics technologies. More recently, it has decided to structure these efforts in the A-Labs, a set of dedicated research facilities including an AI department (AI Lab) focusing more specifically on artificial intelligence. The discussion taking place in this article is related to the activity of this AI Lab, within the A-Labs.

J.-C. Baillie (✉)
Director Aldebaran Robotics AI Lab/A-Labs, Paris, France
e-mail: jcbaillie@novaquark.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_24

24.2 Scientific Background

The long term goal of the research conducted within the AI Lab is centered on the problem of modeling in a robot the full range of cognitive, interactive and social capabilities of humans, within an unconstrained environment, and with a strong emphasis on learning and autonomous interaction.

Many directions have been investigated in the attempt to solve this problem in computational terms for approximately 60 years now. To give a quick and incomplete historical overview of the main trends, we can mention the first pure symbolic approaches (Minsky 1961; Newell et al. 1988; Newell and Simon 1961), which have failed to translate to any useful implementation in robots due to the lack of connection with the action/perception layer. Sub-symbolic approaches have then been advocated to try to reverse the order of importance between cognition and perception and rightfully stress the necessity of embodiment in the development of intelligence (Brooks 1990). However, many of these radical sensori-based approaches, among them neural networks models, have so far failed at scaling up to cognitive levels, beyond reactive or simple task planning models. The recent shift of interest to deep learning and auto-encoding networks (Bengio 2009; Hinton 2007), coupled to ever more powerful CPU capabilities, has merely improved unsupervised clustering capabilities but does not provide any obvious path to a better integrated system design for AI.

In parallel, a paradigmatic shift occurred as some part of the research community progressively realized that the core problem neither lies in the modeling of high level cognitive/symbolic functions of reasoning, nor in the low level reactive perception/action loop, even if both are still challenging today, but in the connection of both in a dynamic and meaningful way, also known as the *symbol grounding problem* (Harnad 1990).

There are two important aspects in the process of grounding symbols into perception. The first one is that grounded symbols are not abstract entities externally imposed to the system by some programmer or expert, but coherent, self-referencing abstractions built by the system itself in reference to a given environment and embodiment. They are meaningful in the sense that they are physically related to what they represent. The second point is that grounded symbols are directly available to be processed by generic cognitive mechanisms to create recursive, hierarchical and structured new symbols that go beyond the direct sensorimotor link. Many results from classical AI can be recruited at this stage, in a field that could be labeled as "grounded cognition" (Barsalou 2008a).

Importantly within this "grounded cognition" framework, the link with non-symbolic sensorimotor roots is preserved and provide the system with sources of reasoning that are not formal but rather experimental. The physical world has regularities and can be simulated, and as such is a non-formal source to induce new rules or inferences of truth about the world, therefore about the symbols themselves.

24.3 Developmental Robotics

A recent research field called *developmental robotics* (Weng et al. 2001; Lungarella et al. 2003; Cangelosi et al. 2010), is trying to address the grounding problem with a particular focus on developmental learning. This idea is an old one, originally expressed by Alan Turing (1950): *"Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain"*.

Unlike most research done in AI and robotics today, which are often limiting the learning to the fine tuning of one particular aspect of a skill or cognitive ability, the developmental program stresses the importance of understanding the dynamical aspect of intelligence, the underlying mechanisms of gradual learning in the context of a progressively more complex interaction with the environment and with other agents (social and environmental grounding).

In developmental robotics, the initial knowledge of the system is limited to a minimal, and if possible unbiased set (for example, in general, no explicit knowledge of the robot body or sensor space is assumed). Scaffolding takes place as new skills are gradually built up on top of previous skills, following an incremental developmental path. It is believed that some steps of this progression are naturally unfolding from internal constraints and dependencies within the system, whereas some could be timed by external signals selected by evolution (Law et al. 2011; Lee et al. 2007; Baranes and Oudeyer 2010).

Unlike some other biologically inspired approaches, in developmental robotics the goal is not directly to model the end product of development, which would be the traditional adult-level intelligence of classical AI, but the process of building up this intelligence itself. The field naturally takes strong inspiration from developmental psychology and cognitive psychology in studying how sensorimotor and cognitive skills are progressively acquired by young children. See Lungarella et al. (2003) and Cangelosi et al. (2010) for a survey of the domain.

The theoretical backbone of developmental robotics is yet to be established beyond design principles, but it can be fruitfully approached through theories of emergence (Jensen 1998; Chialvo 2010), and dynamical systems (Smith and Gasser 2005; Thelen and Smith 1996). The core idea here is that the complexity underlying intelligent behavior and its flexibility cannot be approached in a frontal manner, trying to model the outer part of the observed behavior. It should rather be considered as a irreducible assembly of interdependent grounded cognitive and sensorimotor elements, whose architecture can only be grown and not designed. The scientific challenge is to establish the mechanisms of growth and to identify the constraints (presumably from evolution) that are applied on this growth in the context of a physically grounded system.

One particular way of looking at the problem of intelligence growth is to consider it as the result of the evolution of several coupled dynamical systems, operating

at different time scales and on differents aspects of the body/environment. The decoupling is impossible to do, as all systems can only be explained as the result of their joint interaction, and isolated understanding of one aspect of evolution (for example, language without considering genetics, environmental and social influences, see Steels 2012b) can only be limited.

24.4 Research Directions

The AI Lab conducts research on development and learning within the framework of developmental robotics. The following paragraphs describe the particular problems that we are exploring, but in general all research is more or less directly connected to the problem of language evolution, see Steels (2005), Oudeyer and Kaplan (2006), and Cangelosi and Parisi (2002). We believe that language can play a central role in scaffolding development from low level sensorimotor skills acquisition up to high level cognitive functions, as various aspects of language are relevant to these issues at different developmental stages. We believe also that inspiration from developmental psychology (see for example Tomasello (2005) for a constructivist approach on language acquisition) is particularly important. Note however, that the goal is not necessarily to provide a biologically inspired model for the observed child cognitive development, but rather to take the step by step progression observed in humans as a guideline and potential roadmap for artificial systems.

We detail here the five main research directions that are explored, in increasing degree of abstraction from sensorimotor levels to language acquisition:

- **Low level visual and auditive pattern extraction:** Using recent techniques from deep-learning (Le et al. 2011; Bengio 2009; Behnke 2003) and unsupervised pattern extraction (Wiskott and Sejnowski 2002), we cover the first layer of information and correlation extraction from the environment and proprioception of the robot. The goal is to study what kind of structured information for higher level processing can be grounded from raw sensor data, without introducing bias.
- **Sensorimotor conceptualization:** building on the lowest layers of grounded representations, the task here is to study the possibility to aggregate structures based on mutual information to extract categories (Rakison and Oakes 2003), proto-object definition (Orabona et al. 2007, 2005), simple motor skills schemata (Nishimoto et al. 2008; Mangin and Oudeyer 2012), notions of causality and naive physics (Fitzpatrick et al. 2003). This conceptualization phase will be done in tight integration of perception and action in line with research on active learning (Thrun 1995).
- **Verbal and motor lexicon acquisition through language games:** while the two abstraction layers above could in principle be investigated by an isolated robot interacting with its environment, the next stage involves explicitly social interactions. It is the first layer of language grounding, where the system will have to emerge a shared grounded lexicon for visual and motor categories. It involves

the definition and maintenance of joint attentional scenes and selectionist mechanisms to explore the space of possible symbol conventions. There is a rich literature dealing with these topics, in particular Steels (2012a), Grizou et al. (2012), Barsalou (2008b), Steels and Hild (2012), and Steels (2001)

- **Construction grammar:** the last stage of the developmental program is found in higher cognitive language constructions when disambiguation leads the agents to go beyond lexicon and develop a grammatical structure in language. Difficult computational problems arise as structural matching and unification must be done in a potentially very large space of language constructions. We will follow the constructivist approach of construction grammars, explored among other in particular in Steels (2011), Bergen and Chang (2005), Bates and Goodman (1999), Goldberg (1995), Bybee (2006), and Hirsh-Pasek and Golinkoff (1999). The lexical stage investigated in the lower layers will be used as a way to relate the emergent grammar to sensorimotor perception in an attempt to close the loop of the grounding problem.
- **The Human Cognitive Developmental Map project (CogMap):** as mentioned before, it is of crucial importance to have a good knowledge of the stages of cognitive and language development existing in young children. However, the amount of literature on the topic is extremely large, ever changing and it is usually quite hard to distinguish debated topics from established facts. We propose to launch and coordinate an interdisciplinary open effort to gather current facts about child development inside an online open platform (similar to wikipedia), where established facts and controversies can be debated. The collective effort is crucial as the amount of data to be gathered is far greater that what one individual can accomplish.

More generally, the following directions will be investigated within the different levels of implementation:

- Importance for the system to be able to build *simulations*, or prediction models, of its physical and social environment (theory of mind). This points has been stressed in various context as a key element for cognition (Barsalou 2008a; Hawkins and Blakeslee 2005), affordance modeling (Glenberg and Robertson 2000) or language acquisition (Astington and Baird 2005; Tomasello 1999).
- The relationship between *language and action*, and in particular the possible similitude between higher order recursive representations found in grammar and the same type of compositional constructions found in elaborate non-linear actions plans (see for example references in Cangelosi et al. 2010).
- The role of *pointing* as a key species-specific skill in humans and its relationship with the cognitive and social development of the system. A large literature is devoted to this question, see among others Tomasello (2005), Baron-Cohen (1997), and Scassellati (1999).
- The developmental role of Intrinsic Motivation and system-level goal definition. The balance between exploration and exploitation in the robot interactive patterns (Oudeyer and Kaplan 2006; Schmidhuber 2006).

24.5 Methodology

To help structure the research done in the AI Lab, we have identified several methodological guidelines that play a central role in our efforts:

24.5.1 Integrated Development

There is a vast amount of available literature on developmental robotics covering such topics as grasping, imitation, turn taking, behavior learning, sensorimotor grounding, early language acquisition up to the development of grammar, etc. These researches are extremely useful because they allow to test and validate hypothesis in isolated experiments. However, as we progressively gain more understanding at several levels of the developmental progression, it becomes also necessary to try and integrate systems together. This effort will reveal other types of problems linked precisely at the interface of different algorithms, help progress towards a more generalized theoretical understanding of the underlying mechanisms, and can foster cooperation between several teams. This integration constraint has several consequences:

1. Generalized data structures or models must be defined at the level of the whole system, to enable components to interface each other.
2. When work is done at a certain level of abstraction (for example, working on grammar acquisition), it might be necessary to emulate the underlying layers which are themselves in development and not yet ready. This must be done with a very clear view of the perspective of replacing the emulation by the output of the other layers.
3. Beyond the software level of the integration constraint, collaborative team work is crucial to ensure a focused effort.

24.5.2 Continuous Learning

Most interesting developmental scenario requires a certain amount of interaction time between the robot and an operator, or between the robot and its environment. This is needed to create a rich set of perception and action opportunities from which learning is possible. Besides, as we position our research in a developmental framework, it is important to imagine a continuous gradual acquisition of representations and abstractions by the robot, which will feed each other at several levels as the learning keeps going. New skills can help to build other skills, in a virtuous circle. This approach is known as *life long learning* (Thrun 1994).

24.5.3 Autonomous Behavior and Intrinsic Motivation

The term *autonomous* denotes the fact that the behavior of the robot will not be dictated by predefined tasks or goals, but should emerge out of more basic principles governing the robot actions. This is usually referred as *Intrinsic Motivation* or artificial curiosity, see for example Oudeyer et al. (2005), Oudeyer and Kaplan (2006), and Baranes and Oudeyer (2013).

While the behavior of the robot is autonomous in the above sense, it will alternate between phases of (weakly) supervised and unsupervised learning. Supervised learning phases require the implementation of an experimental protocol where non-expert users are invited to interact with the robot on a daily basis.

24.5.4 Robotics Experimentation

Our experiments are run on real robots, excluding simulation. In particular, it is difficult to provide a good simulation for human-robot interaction, especially in the context of interaction with naive users.

24.6 Conclusion

The AI Lab of Aldebaran Robotics is a relatively unique structure for research, focusing on fundamental Science within the framework of developmental robotics. The purpose of the lab is to advance the company's understanding on these core problems. The expected results can be defined around three global milestones ranging from short to long term:

Short term: better sensor fusion and integration of sensorimotor information in a coherent and grounded picture of the robot environment. First proto categories of sensorimotor events.

Medium term: first set of language games available (lexicon level), creation of shared grounded symbolic convention between a robot and a user, emergence of multi-modal non-verbal communication (pointing, gestures, sounds).

Long term: full fledged grammar-capable system that can be taught to learn a language including simple grammatical structures, and compositional actions in a natural interactive way.

Overall, applications to commercial products are envisioned at different stages of progress, from low-level more robust perceptual strategies, up to interaction modules and smooth learning capabilities to evolve the robot and improve its capability to adapt to the user needs. We believe that superficial reproduction of social behaviors in robots can only reach a plateau of performance and users will be quick to detect the non-grounded underlying mechanisms. The preferred long term

direction should favor genuine social grounding and human-robot shared communication/interaction emergence. This will improve the acceptability and usefulness of robots in various settings, opening doors to numerous fruitful applications to help people in their daily lifes.

References

Astington, J. W., & Baird, J. A. (2005). *Why language matters for theory of mind*. Oxford: Oxford University Press.

Baranes, A., & Oudeyer, P. Y. (2010). Maturationally-constrained competence-based intrinsically motivated learning. In *IEEE 9th International Conference on Development and Learning (ICDL), 2010*, Ann Arbor (pp. 197–203). IEEE.

Barnes, A., & Oudeyer, P. Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems, 61*(1), 49–73.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge: MIT.

Barsalou, L. (2008a). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

Barsalou, L. (2008b). Grounding symbolic operations in the brain's modal systems. In *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. Cambridge/New York: Cambridge University Press.

Bates, E., & Goodman, J. (1999). On the emergence of grammar from the lexicon. *The emergence of language* (pp. 29–79). Mahwah: Lawrence Erlbaum Associates.

Behnke, S. (2003). *Hierarchical neural networks for image interpretation* (Vol. 2766). Berlin/New York: Springer.

Bengio, Y. (2009). Learning deep architectures for ai. *Foundations and Trends® in Machine Learning, 2*(1), 1–127.

Bergen, B., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In *Construction grammars: Cognitive grounding and theoretical extensions* (pp. 147–190). Amsterdam/Philadelphia: John Benjamins.

Brooks, R. (1990). Elephants don't play chess. *Robotics and Autonomous Systems, 6*(1), 3–15.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*, 711–733.

Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language*. London: Springer

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., & Nori, F., et al. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development, 2*(3), 167–195.

Chialvo, D. R. (2010). Emergent complex neural dynamics. *Nature Physics, 6*(10), 744–750.

Fitzpatrick, P., Metta, G., Fitzpatrick, P., & Metta, G. (2003). Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences, 361*(1811), 2165–2185.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*(3), 379–401.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Grizou, J., Lopes, M., & Oudeyer, P. (2012). Robot learning simultaneously a task and how to interpret teaching signals. In *IEEE-RAS International Conference on Humanoid Robots*, Madrid.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1), 335–346.

Hawkins, J., & Blakeslee, S. (2005). *On intelligence*. New York: St. Martin's Griffin.
Hinton, G. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*(10), 428–434.
Hirsh-Pasek, K., & Golinkoff, R. M. (1999). *The origins of grammar: Evidence from early language comprehension*. Cambridge: MIT.
Jensen, H. J. (1998). *Self-organized criticality: Emergent complex behavior in physical and biological systems* (Vol. 10). Cambridge/New York: Cambridge university press.
Law, J., Lee, M., Hülse, M., & Tomassetti, A. (2011). The infant development timeline and its application to robot shaping. *Adaptive Behavior, 19*(5), 335–358.
Le, Q. V., Monga, R., Devin, M., Corrado, G., Chen, K., Ranzato, M., Dean, J., & Ng, A. Y. (2011, preprint). Building high-level features using large scale unsupervised learning. arXiv:11126209.
Lee, M. H., Meng, Q., & Chao, F. (2007). Staged competence learning in developmental robotics. *Adaptive Behavior, 15*(3), 241–255.
Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: A survey. *Connection Science, 15*(4), 151–190.
Mangin, O., & Oudeyer, P. (2012). Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization. In *RSJ International Conference on Intelligent Robots and Systems*, Vilamoura-Algarve.
Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8–30.
Newell, A., Shaw, J., & Simon, H. A. (1988). Chess-playing programs and the problem of complexity. In *Computer games I* (pp. 89–115). New York: Springer.
Newell, A., & Simon, H. A. (1961). *Computer simulation of human thinking*. Santa Monica: Rand Corporation.
Nishimoto, R., Namikawa, J., & Tani, J. (2008). Learning multiple goal-directed actions through self-organization of a dynamic neural network model: A humanoid robot experiment. *Adaptive Behavior, 16*(2–3), 166–181.
Orabona, F., Metta, G., & Sandini, G. (2005). Object-based visual attention: A model for a behaving robot. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops*, San Diego (pp. 89–89). IEEE.
Orabona, F., Metta, G., & Sandini, G. (2007). A proto-object based visual attention model. In *Attention in cognitive systems: Theories and systems from an interdisciplinary viewpoint* (pp. 198–215). Berlin/Heidelberg: Springer.
Oudeyer, P. Y., & Kaplan, F. (2006). Discovering communication. *Connection Science, 18*(2), 189–206.
Oudeyer, P., Kaplan, F., Hafner, V., & Whyte, A. (2005). The playground experiment: Task-independent development of a curious robot. In *Proceedings of the AAAI Spring Symposium on Developmental Robotics*, Stanford (pp. 42–47).
Rakison, D., & Oakes, L. (2003). *Early category and concept development: Making sense of the blooming, buzzing confusion*. Oxford/New York: Oxford University Press.
Scassellati, B. (1999). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *Computation for metaphors, analogy, and agents* (pp. 176–195). Berlin/New York: Springer.
Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science, 18*(2), 173–187.
Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life, 11*(1–2), 13–29.
Steels, L. (2001). Language games for autonomous robots. *Intelligent Systems, IEEE, 16*(5), 16–22.
Steels, L. (2005). The emergence and evolution of linguistic structure: From lexical to grammatical communication systems. *Connection Science, 17*(3–4), 213–230.
Steels, L. (2011). *Design patterns in fluid construction grammar* (Vol. 11). Amsterdam/Philadelphia: John Benjamins.

Steels, L. (2012a). *Experiments in cultural language evolution* (Vol. 3). Amsterdam/Philadelphia: John Benjamins.

Steels, L. (2012b). Interactions between cultural, social and biological explanations for language evolution. *Physics of Life Reviews, 9*(1), 5–8.

Steels, L., & Hild, M. (2012). *Language grounding in robots*. New York: Springer.

Thelen, E., & Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action*. Cambridge: MIT.

Thrun, S. (1994). A lifelong learning perspective for mobile robot control. In *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems'94.'Advanced Robotic Systems and the Real World', IROS'94* (Vol. 1, pp. 23–30). Munich: IEEE.

Thrun, S. (1995). Exploration in active learning. In M. Arbib (Ed.), *Handbook of brain. Science and neural networks* (pp. 381–384). Cambridge, MA: MIT Press.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge: Harvard University Press.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen, E. (2001). Autonomous mental development by robots and animals. *Science, 291*(5504), 599–600.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715–770.

Chapter 25
Tacit Representations and Artificial Intelligence: Hidden Lessons from an Embodied Perspective on Cognition

Elena Spitzer

Abstract In this paper, I explore how an embodied perspective on cognition might inform research on artificial intelligence. Many embodied cognition theorists object to the central role that representations play on the traditional view of cognition. Based on these objections, it may seem that the lesson from embodied cognition is that AI should abandon representation as a central component of intelligence. However, I argue that the lesson from embodied cognition is actually that AI research should shift its focus from how to utilize explicit representations to how to create and use tacit representations. To develop this suggestion, I provide an overview of the commitments of the classical view and distinguish three critiques of the role that representations play in that view. I provide further exploration and defense of Daniel Dennett's distinction between explicit and tacit representations. I argue that we should understand the embodied cognition approach using a framework that includes tacit representations. Given this perspective, I will explore some AI research areas that may be recommended by an embodied perspective on cognition.

Keywords Embodied cognition • Artificial intelligence • Representation • Tacit representations

25.1 Introduction

In the past 20 years, embodied cognition (EC) has emerged as a proposed alternative or enhancement to traditional cognitive science. Embodied cognition can be seen as a research program that encourages us to pay more attention to the role that the rest of the body, not just the brain, plays in cognition. One common EC objection against traditional cognitive science has to do with the role that internal representations are presumed to play in cognition. When we pay more careful attention to the role of

E. Spitzer (✉)
University of Wisconsin, Madison, WI, USA
e-mail: emspitzer1@gmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_25

the body in cognition, it often seems that the types of representations postulated by traditional cognitive science are explanatorily inadequate.

In this essay, I consider what lessons an embodied perspective on cognition might hold for the development of artificial intelligence. Artificial intelligence had ambitious beginnings in the latter part of the twentieth century, with grand predictions of human–level intelligence right around the corner. However, AI research has since run into a number of difficult challenges, and the original AI dream of general intelligence has faded. Current research on and applications of AI (especially the most successful ones) focus on "specialized intelligence." In other words, the focus is on ways to make computers better at the sorts of tasks that humans are bad at—such as massive information storage and fast retrieval, very large scale data analysis, matching advertisers with search engine users, etc. There is less focus on "general intelligence"—how to build a computer that demonstrates some of the sorts of intelligence that come so naturally to humans. Progress towards building a computer that is able to do tasks that a five year old could perform easily, like playing hide-and-go-seek, has been slow.

In this paper, I will explore how an embodied perspective on cognition might inform research on developing general intelligence in AI. I will discuss arguments from three embodied cognition theorists—Andy Clark, Hubert Dreyfus, and Rodney Brooks. As critics of the classical view of cognition, each of these authors objects to the central role that representations play on the traditional view. Based on these arguments, it may seem as though the lesson from embodied cognition is that AI should abandon representation as a central component of intelligence.

However, I will argue that the lesson from embodied cognition is actually that AI research should shift its focus from how to utilize *explicit* representations to how to create and use *tacit* representations. In Sect. 25.1, I will provide an overview of the commitments of the classical view of cognition. Section 25.2 describes the role of representations in the classical view and three different critiques of that role. In Sect. 25.3, I introduce Daniel Dennett's distinction between three types of representation (explicit, implicit, and tacit), and provide further exploration and defense of tacit representations. The final section demonstrates how the embodied cognition perspective from Clark, Dreyfus, and Brooks could be helpfully reconceived in a framework that includes tacit representations. In this final section I will also explore some AI research areas that may be recommended by an embodied perspective on cognition.

25.2 The Classical View

The standard view in cognitive science is that cognition consists of a certain type of computational process. More specifically, the brain is presumed to be some kind of symbol system, and cognition proceeds by symbol manipulation. Newell and Simon state the Physical Symbol System Hypothesis (PSSH) as follows:

> *The Physical Symbol System Hypothesis*. A physical symbol system has the necessary and sufficient means for general intelligent action. (Newell and Simon 1976, p. 292)

The PSSH is taken to be an empirical hypothesis, and some version of this hypothesis has formed the basis of most theorizing and research in cognitive science. On the basis of the PSSH, cognitive science set out to locate the cognitive symbols and the rules that govern their manipulation.

The idea that the brain is a physical symbol system generally involves (at least) the following commitments. First, physical parts of the brain act as symbols in the cognitive system. For something to be a symbol means that it represents something else (i.e. has some semantic content), that its relationship to what it represents is arbitrary (e.g. the word "dog" does not look like a dog), and that it can potentially be combined with other symbols to create new symbols (Shapiro 2010, pp. 10–11).

Additionally, the PSSH includes the notion that thinking involves the process of symbol manipulation. For example, the cognitive process of reasoning from the proposition "Socrates is short and Socrates is ugly" to the proposition "Socrates is short" is analogous to the standard inference rule of conjunction elimination in first-order logic. There is some part of the brain that serves as a symbol for "Socrates is short" and another symbol for "Socrates is ugly." In sentential logic, we derive "S" from "S & U" based on the syntactic structure of the conjunction of the two symbols "S" and "U" and the rule of conjunction elimination. Analogously, according to the PSSH, the brain reasons from "Socrates is short and Socrates is ugly" to "Socrates is short" according to the syntactic structure of the proposition and rules about reasoning.

One particularly influential elaboration on the PSSH is Jerry Fodor's Language of Thought Hypothesis. The Stanford Encyclopedia describes this hypothesis as follows:

> The Language of Thought Hypothesis (LOTH) postulates that thought and thinking take place in a mental language. This language consists of a system of representations that is physically realized in the brain of thinkers and has a combinatorial syntax (and semantics) such that operations on representations are causally sensitive only to the syntactic properties of representations. According to LOTH, thought is, roughly, the tokening of a representation that has a syntactic (constituent) structure with an appropriate semantics. Thinking thus consists in syntactic operations defined over such representations. (Aydede 2010)

The standard view holds that cognition proceeds via the algorithmic manipulation of syntactically structured explicit representations in the brain.

25.3 Embodied Cognition Critiques

A number of criticisms of the classical view in cognitive science and its attendant research programs in AI have focused on the role that representations are presumed to play in cognition—human or artificial.

I will discuss three different proponents of embodied cognition who have criticized the role of representations in the classical view in one way or another—Andy Clark, Hubert Dreyfus, and Rodney Brooks. Broadly speaking, embodied cognition theorists argue that the classical view in cognitive science is *too cerebral* in some sense. Specifically, they think that the classical view does not take sufficient account of the role that the body (other than the brain) plays in cognition.

The classical view presumes that the body's role in cognition is fairly limited. The body serves as an input and output device—delivering signals to the brain and carrying out its commands. It is supposed that the truly cognitive processes consist only of what happens between the input and output. Shapiro puts the point as follows: "All the 'action,' so to speak, begins and ends where the computational processes touch the world. The cause of the inputs and the effect on the world that the outputs produce are, from this perspective, irrelevant for purposes of understanding the computational processes taking place in the space between input and output" (Shapiro 2010, p. 26).

Against the classical view, proponents of embodied cognition claim that the body often plays an important role in cognition, which the classical view fails to recognize. Thus, EC proponents maintain that cognitive science must take better account of cognitive agents' bodies in order to understand cognition more fully. The evidence for this line of thinking is varied, but the message is similar.

25.3.1 The Role of Representations

Representations play a particularly important role in the classical view. For example, Fodor and Pylyshyn characterize the classical view in terms of (1) the properties of mental representations and (2) the properties of the processes that transform those mental representations (Fodor and Pylyshyn 1988, pp. 12–13). Representations are the symbols part of a physical symbol system. And they are central to the Language of Thought Hypothesis, according to which internal representations are the components that make up the language of thought.

Because of this prominent role, representations are often a target for objections to the classical view. Arguments against the classical view from proponents of embodied cognition often follow this general pattern: The body plays such-and-such a role in cognition. Representations are the only means of explaining cognitive activity on the classical view. But representations cannot explain the role that the body plays in so-and-so. Therefore, the classical view is insufficient for explaining cognition.

Below, I will discuss three arguments from EC proponents that target the role of representations in the classical view. Each of the arguments finds that representations are inadequate for various reasons, and thus concludes that the classical view must be amended, augmented, or perhaps even abandoned.

25.3.2 Clark's Cricket

Internal representations are a necessary component of cognition on the classical view. However, they often seem ill-suited to the more embedded and embodied components of cognition that EC theorists focus on. Some EC theorists advocate abandoning internal representations entirely. Clark does not agree that we can do away with internal representations altogether, but he does think it may be appropriate to deemphasize internal representations in many areas of cognition.

Clark discusses the following example from Barbara Webb's research as a case in which a seemingly intellectual and representation-dependent task is better understood as a mechanistic and non-symbolic task.

> Phonotaxis is the process whereby a female cricket identifies a male of the same species by his song, turns in his direction, and reliably locomotes to the source. On the face of it, this looks to be a complex, three part problem. First, hear various songs and identify the specific song of the male of your own species. Second, localize the source of the song. Third, locomote that way (making needed corrections en route). This way of posing the problem also makes it look as if what the cricket will require to solve the problem are some quite general cognitive capacities (to recognize a sound, to discern it's source, to plot a route). Nature, however is much thriftier, and has gifted the cricket with a single, efficient, but totally special-purpose problem-solving procedure . . .
>
> This whole system works only because first, the cricket's tracheal tube is especially designed to transmit songs of the species-specific frequency, and because the male song repeats, offering frequent bursts each of which powers one of these episodes of orienting-and-locomoting (hence allowing course corrections, and longer distance mate-finding) . . .
>
> In the very specific environmental and bio-mechanical matrix that Webb describes, the cricket thus solves the whole (apparently three-part) problem using a single special–purpose system. There is no need, for example, to . . . bother to build a model of your local surroundings so as to plan a route. Instead, you (the cricket) exploit neat tricks, heuristics, and features of your body and world. Moreover you (and your realworld sisters) seem to succeed *without relying on anything really worth calling internal representations*. (Clark 2001, pp. 127–128 emphasis added)

There is no agreed upon definition of what makes an activity cognitive. But for the sake of argument, I think we can treat phonotaxis as cognitive here. Though I think there are other good reasons to count phonotaxis as cognitive. One reason is that this type of flexible interaction with the world seems to be essential for general intelligence. And, insofar as my project here concerns general artificial intelligence, we need some account of how this type of activity works. Especially since this is exactly the kind of thing that computers are much worse at than living organisms. Rodney Brooks, a roboticist that we'll meet again later in the paper, stresses the importance of this type of "low-level" activity: "Our goal . . . is simple insect level intelligence within two years. Evolution took 3 billion years to get from single cells to insects, and only another 500 million years from there to humans. This statement is not intended as a prediction of our future performance, but rather to indicate the nontrivial nature of insect level intelligence" (Brooks 1991, p. 156). Being able to explain activities such as phonotaxis is likely an important step in better understanding cognition and developing general AI. So, even if we do not

want to call phonotaxis cognitive *per se*, it does seem to be the type of activity that is a precondition for "real" cognitive activity—and that's cognitive enough for my purposes here.

According to the classical view, the most natural way to break down the cricket's behavior is as a three step process—recognize, locate, and move towards the song. On this model, the female cricket has internal representations of what she hears, of whether it matches the song in which she is interested, of where the sound is coming from, how to get there, etc. But the way that phonotaxis actually happens does not seem to mirror this structure at all. The observed methods seem to have less in common with rule-based inference and more in common with simply mechanical operations—no rules and no representations, according to Clark.

Clark suggests that the sort of mechanisms responsible for cricket phonotaxis may be very common cognitive mechanisms in humans as well. But this requires that the classical view either be supplemented or replaced.

25.3.3 Dreyfus & Dreyfus and the Foundations of AI

Another philosopher who has criticized the role of representations in artificial intelligence is Hubert Dreyfus. Hubert Dreyfus and his brother, Stuart Dreyfus, provided an early and influential critique of artificial intelligence.

Representations play a central role in the Dreyfus & Dreyfus critique of the classical view. The quick, but hopefully not too dirty version of the argument goes like this: AI is doomed because genuine intelligence requires a commonsense background that computers can never possess. Dreyfus says, "Intelligence requires understanding, and understanding requires giving a computer the background of common sense that adult human beings have by virtue of having bodies, interacting skillfully with the material world, and being trained into a culture" (Dreyfus 1992, p. 3). The type of commonsense background that Dreyfus has in mind can be demonstrated by his example of the "cultural *savoir faire*" of gift-giving—the ease with which an adult human in a particular culture can give an appropriate gift at an appropriate time and in an appropriate way (Dreyfus 1992, p. xxiii). For example, when you are walking through a shop looking for a gift to take to a friend's dinner party, you may stop to consider a bouquet of flowers, but you do not stop to consider bringing 20 cans of tuna fish or a box of roach motels.

Intelligence requires a commonsense background, but Dreyfus & Dreyfus think this background cannot be represented. They believe that this background consists of a set of skills, not a body of knowledge—i.e. commonsense background requires knowledge-how, not knowledge-that. However, Dreyfus & Dreyfus believe this is the sort of information that cannot be represented by computers. They say, "If background understanding is indeed a skill and if skills are based on whole patterns and not on rules, we would expect symbolic representations to fail to capture

our commonsense understanding" (Dreyfus and Dreyfus 1988, p. 33). Dreyfus & Dreyfus think that the commonsense background cannot be enumerated and represented as a set of facts, and that this is what would be required in order to get a computer to act intelligently. Thus, because of their reliance on representations, the classical view will be incapable of fully explaining intelligence and AI will be incapable of building genuine intelligence.

25.3.4 Brooks' Creatures

Rodney Brooks is a roboticist from MIT. In his article "Intelligence Without Representation," he describes some of his work building intelligent autonomous mobile robots, which he calls "creatures." Brooks' creatures were groundbreaking because they were relatively good at performing tasks in real-world environments. Previous robots could perform in highly structured environments, but could not deal with the variability of a real world environment.

Brooks attributes his creatures' success in real-world environments to his unique approach to their design—an approach that does not use representations as traditionally conceived. Previous approaches to mobile robot design had the robot create representations of the world based on information gained from its sensors. These robots would then have a central processor reason about those representations in conjunction with its goals (also represented) in order to decide what to do next.

Brooks calls his design approach the "subsumption architecture." On this approach, instead of having external sensors and a centralized internal processor, there are multiple layers of control, each of which interacts with the world and "makes decisions" on its own.

Like Clark and the Dreyfus brothers, Brooks also wants to diminish or eliminate the role of representation in cognition. Indeed, the title of his 1991 paper is "Intelligence Without Representation." Brooks attributes the success of his creatures to the absence of representations. He says, "a common theme in the ways in which our layered and distributed approach helps our Creatures meet our goals is that there is no central representation" (Brooks 1991, p. 147).

Brooks' work demonstrates that "there need be no explicit representation of either the world or the intentions of the system to generate intelligent behaviors for a Creature" (Brooks 1991, p. 149). This does not show that intelligent behavior never requires explicit representations, but it certainly seems to demonstrate a more promising approach to realworld interaction than previous attempts. And, if Brooks is right that being able to act in a complex and dynamic environment is a cornerstone of intelligence, then this is potentially a more promising approach for AI than the one more informed by the classical view.

25.3.5 Ostensible Lesson for AI

All of these theorists take issue with the role that representations have played in standard accounts of cognition. Clark argues that the classical view is too quick to posit representations, when apparently non-representational mechanisms might account for cognitive phenomena. Dreyfus and Dreyfus argue that genuine intelligence requires knowhow and that this knowhow cannot be represented. For them, this implies that computers are not capable of intelligence because computers are limited to using information that can be represented. Finally, Brooks argues that representations are the wrong way to proceed in order to build robots that can interact effectively with a dynamic environment. He eschews representations in favor of an alternative design strategy that he claims does not rely on representations at all.

All three viewpoints seem to suggest that the lesson for AI from an embodied perspective on cognition is that if AI wants to achieve general intelligence, then AI theorists should shift their focus away from representations. The lesson seems to be that representations are the wrong way to go in order to achieve general intelligence of the sort displayed by humans (or other animals). In varying ways and to varying degrees, all three authors suggest that representations are inadequate to account for the kind of skilled interaction that is necessary for general intelligence.

This *seems* to be the lesson for AI from an embodied perspective on cognition, but I think it is problematic. I think this lesson is problematic because any intelligent system must have some way of obtaining and using information about its environment. And anything that serves the role of conveying information is a representation of some sort. So obtaining and using information about one's environment actually requires representations. Thus, advising AI theorists to abandon representations is incoherent if it amounts to a suggestion that they should give up trying to build artificial systems that obtain and use information about the environment. I will revisit this idea again below.

I will argue that there is actually a different lesson for AI from an embodied perspective on cognition. Instead of advising AI to abandon representations, I think the lesson is that AI theorists should focus on a particular kind of representation—tacit representations. I will explain what tacit representations are and how they relate to embodied cognition and AI in the next two sections.

25.4 Are All Representations Created Equal?

Each author discussed above thinks that at least some cognition does not involve representations at all. Their work also suggests the even stronger claim that non-representational cognitive mechanisms are a necessary feature of intelligence. Below I will argue that we can develop an alternative to the classical view without giving up on representations altogether.

25.4.1 Basics of Representation

In order to proceed, we first need to have some working definition of representation. There is no agreed upon definition, but Fred Dretske has developed a plausible account, and I'll use that as my working model here. Prinz and Barsalou (2000) adopt this approach as well and provide the following useful summary of Dretske's views:

> Many philosophers believe that representation involves information. A state represents a property only if it carries information about it (Dretske 1981; Fodor 1990). Carrying information is, in turn, analyzed in terms of nomic, or law-like, covariation. To a first approximation, a state s carries information about a property F just in case instantiations of F reliably cause tokens of s. Although arguably necessary, information is not sufficient for representation. Fire reliably causes smoke, but smoke does not represent fire. An information-bearing state must satisfy one further condition to count as a representation. For Dretske (1995), this extra ingredient is teleological: Something can represent something else only if it has the function of carrying information about it. Representations can acquire such functions through a variety of different histories, including natural selection, learning, and design. (Prinz and Barsalou 2000, p. 55)

So, there are two components of representations on Dretske's account—an information component and a teleological component. Below, I will outline three types of representation and use Dretske's account to assess one of them.

25.4.2 Three Styles of Representation

In "Styles of Mental Representation," Daniel Dennett distinguishes between three types of representation: explicit, implicit, and tacit. Dennett wants to explore the ways in which computers represent in order to investigate the analogous features in human cognition.

Dennett suggests that cognitive science has been overly focused on explicit representations. Explicit representations are the most obvious kind of representations in a computer because these are what we deal with when we program computers. Perhaps as a result of this, the classical approach to cognitive science has focused on analogous theoretical structures in the human mind. However, Dennett argues that computers also contain tacit representations, and that these might be more informative and useful as a theoretical tool in cognitive science than has previously been appreciated.

Explicit representations are the most familiar type of representation. The word "cat" explicitly represents a cat, the word "bicycle" explicitly represents a bicycle, and the picture below (Herford 1899) explicitly represents a cat riding a bicycle.

Each of these is an explicit representation because it is a physical object that has semantic content according to some system of interpretation. An explicit representation is physically stored in a system and can be interpreted and used by the system (Dennett 1982).

Dennett defines an implicit representation as one that is logically implied by explicit representations in the system. For example, if a system explicitly represents "Eeyore is a donkey" and "All donkeys have ears," then it implicitly represents "Eeyore has ears." Thus, you could (implicitly) represent the proposition "Eeyore has ears," even if you had never considered the matter of Eeyore's ears before.

Dennett distinguishes a third type of representation—tacit representations. Dennett's description of tacit representations is less well–developed than his description of either explicit or implicit representations. He uses the example of a pocket calculator to clarify the idea. The example starts by asking, "Does a pocket calculator represent the 'truths of arithmetic' explicitly, implicitly, or tacitly?" (Dennett 1982, p. 221).

The calculator produces correct answers to arithmetical queries without explicitly referencing any rules of arithmetic. When the calculator computes "$2 + 2$," it does not produce "4" by looking up a rule that says "$2 + 2 = 4$." Dennett says, "the calculator is a device with the dispositional competence to produce explicit answers to explicit questions... but it does this without relying on any explicit representations within it—except the representations of the questions and answers that occur at its input and output edges and a variety of interim results" (Dennett 1982, p. 222). The calculator's circuitry is designed such that it performs addition according to the rules of arithmetic, but these rules are not themselves explicitly represented in the system. The rule that "$2 + 2 = 4$" is not stored in the system

and recalled and used when necessary. Nor are the rules of arithmetic implicitly represented by the calculator. Instead, engineers built the calculator so that it would operate in a way that conforms to the rules of arithmetic. Thus, we might say that the calculator tacitly represents how to do arithmetic.

As a first attempt to generalize from this example, perhaps we can say that a representation is tacit when it is, in some sense, *built into* a system. An explicit representation is something that the system *uses as* a representation in order to accomplish some task. For example, if we had a calculator that worked by using a lookup table, then the entries in the look up table would be explicit representations because of how they are used by the system. By contrast, although a tacit representation also has a physical presence in the system, it is not used *by* the system. Instead, a tacit representation is the thing in virtue of which the system displays some disposition.

The calculator's circuitry makes it possible for the calculator to do addition—to have the disposition to produce the correct input-output pairs. But the calculator does not consult its circuitry in order to determine how to perform addition. Instead, the calculator's particular circuitry is the mechanism in virtue of which it is able to perform addition. The rules of arithmetic are not stored and consulted. Nor does the circuitry itself consult any rules. The rules of arithmetic are the reason that the circuitry does what it does—since they guided its design. The calculator has arithmetic knowhow because its circuitry tacitly represents how to do arithmetic.

25.4.3 A Case for Tacit Representations as Representations

Some might object that knowhow is not the kind of thing that can be represented. Propositions can be represented but skills and dispositions cannot. However, I think there are good reasons to consider tacit representations as *bona fide* representations.

According to Dretske's account, representations comprise both an information component and a teleological component. Below I will argue that tacit representations include both components.

The information criterion says that state S represents Q only if S carries information about Q. In the case of tacit representations, I propose that a system that tacitly represents how to Q contains information about how to Q. The nature information is a debated topic (Adriaans 2012; Floridi 2013), and I do not propose to attempt to settle the matter here. And so, in order to make use of Dretske's information criterion, I will appeal to an everyday (though admittedly vague) understanding of the word.[1]

[1]This is Dretske's strategy as well. For example, in "The Epistemology of Belief," Dretske says, "I have begun to talk more and more about information, so let me pause a moment to explain what I mean by this way of talking. I mean nothing very technical or abstract. In fact, I mean pretty much what (I think) we all mean in talking of some event, signal or structure carrying (or embodying) information about another state of affairs. A message (i.e., some event, stimulus or signal) carries

On this everyday understanding, I think it is reasonable to say that knowhow involves using information about how to do something. For example, when someone knows how to do a flip-turn while swimming, they have information about how to do a flip-turn. One must learn how to do a flip-turn, and I think this can plausibly be understood (at least in part) as gaining information about how to do one. Knowhow requires information, and I propose that tacit representations carry that information.

According to the discussion above, a tacit representation is the structure(s) in virtue of which a system performs according to its knowhow. Given this definition of a tacit representations and the idea that knowhow has informational content, I think we must recognize tacit representations as carrying information about knowhow. The pocket calculator is able to do arithmetic, and its circuitry contains information about how to do arithmetic. My "circuitry" contains information about how to do a flip-turn, and so I know how to do a flip-turn. Thus, tacit representations meet Dretske's information criterion.

Dretske's second criterion is the teleological criterion. The teleological criterion requires that representations *have the function* of representing the thing which they represent.

I think tacit representations meet this criterion as well. Things acquire functions in different ways, so we may need to judge on a case-by-case basis. But there is nothing about tacit representations in and of themselves that would seem to violate this condition. And certainly in the case of the calculator's circuitry, by design, it has the function of carrying information about how to do arithmetic.

Thus, it is not the case that knowledge-that can be represented while knowledge-how cannot. Both can be represented in a system. But the role that each plays in the system is quite different. Knowledge-that is explicitly or implicitly represented, while knowledge-how is tacitly represented.

25.5 Lessons for AI

25.5.1 A False Dilemma

Embodied cognition proponents often seem to suggest that we must choose between representation and embodied cognition. However, if we employ a framework that includes tacit representations, then this is turns out to be a false dilemma. I think we need not choose between representations and an embodied approach to cognition.

The classical view says that cognition involves the manipulation of explicit representations on the basis of syntactic rules. All three of the embodied cognition theorists discussed above reject this view because it cannot account for various important cognitive phenomena. The notion of internal representations is clearly an

information about X to the extent to which one could learn (come to know) something about X from the message" (Dretske 1983, p. 10).

important component of the classical view. And, I think these embodied cognition theorists are right to contend that the role representations play in the classical view is particularly ill-suited to account for certain important cognitive phenomena. But that does not mean that representations themselves are not a part of the explanation for these phenomena. Instead, it only implicates the particular commitments that the classical view has concerning representations—namely that explicit representations should be used to explain cognitive phenomena.

But tacit representations are not faulted in the same way as explicit representations by these arguments. In fact, I think a framework that includes tacit representations can more easily accommodate the examples that Clark, Brooks, and Dreyfus discuss. I will revisit each of these author's examples below to demonstrate how the examples would fit into the tacit representation framework.

25.5.2 Crickets, Creatures, and Commonsense

25.5.2.1 Crickets

Clark describes the crickets' phonotaxis machinery as a case where there is no representation of the world or the rules that guide the crickets' behavior. Instead, he says, the crickets "seem to succeed without relying on anything really worth calling internal representations" (Clark 2001, p. 128). Clark does not tell us what makes something worthy or unworthy of the name "internal representation," but I think it's safe to assume he has explicit representations in mind.

I think Clark is right to judge that explicit representations are out of place in explaining cricket phonotaxis. However, I think it would be quite reasonable to understand the cricket's phonotaxis machinery as a tacit representation. Features of the cricket's bodily setup tacitly represent its mate-detection-knowhow. There are a lot of details included in this knowhow—how to recognize a potential mate's song, how to locate him, how to get there, and a multitude of other elements. But, as Clark argues, it is the cricket's physical features in virtue of which it is able to perform this task. This seems to fit quite well into the tacit representation framework discussed above. Unlike the calculator's circuitry, the cricket's circuitry was not *designed* to function in this manner. But it sill has an appropriate teleology—since presumably it was *selected for* in order to perform this function.

25.5.2.2 Creatures

Brooks offers a sentiment that is similar to Clark's about what's "worth" calling a representation:

> Even at a local level we do not have traditional AI representations. We never use tokens which have any semantics that can be attached to them... An extremist might say that we really do have representations, but that they are just implicit... *However we are not happy*

> *with calling such things a representation. They differ from standard representations in too many ways* . . . There are no choices to be made. To a large extent the state of the world determines the action of the Creature. (Brooks 1991, p. 149, emphasis added)

Brooks occasionally refers to the representations he's rejecting as "explicit representations," but he's not consistent in his usage. He often seems to suggest that his creatures do not use any representations at all.

I think a representational framework that includes tacit representations is compatible with and even complements Brooks' subsumption architecture idea. Within this framework, Brooks' focus on system architecture design can be understood as a technique to implement knowhow in a system. The subsumption architecture enables the creatures to perform skillfully in their environment without having to represent facts about the environment and reason about them.

Brooks suggests that the fact that the state of the world (directly) determines the action of his creatures is a reason in favor of rejecting the idea that there are representations involved. Instead of representing the world and then using rules to reason about those representations and decide on an action, Brooks' creatures are designed so that they interact with the world in a more direct fashion. This is telling against explicit representations, but not so for tacit representations. Instead, I think tacit representations are quite suitable for describing Brooks' subsumption architecture.

25.5.2.3 Commonsense

In *What Computers Still Can't Do*, Hubert Dreyfus elaborates on his objection to AI and representations. He discusses the rationalist history of AI—philosophers such as Descartes and Leibniz thought the mind could be defined by its capacity to form representations of the "fixed" and "context-free" features of a domain and the rules governing their interactions. According to Dreyfus, on this view everything that we know, including knowhow, must be stored in the mind in propositional form. He calls this view of the mind "representationalism" (Dreyfus 1992, p. xvii).

Dreyfus argues that genuine intelligence requires understanding and understanding requires a rich background that can only be obtained by having a body and interacting constantly and skillfully with the material and cultural world (Dreyfus 1992, p. 3). With representationalism as the foundation, traditional AI assumes that "common sense derives from *a vast data base of propositional knowledge*" (Dreyfus 1992, p. xvii). The "common sense problem" for AI lies in how to go about creating, structuring, and querying this database so as to create genuinely intelligent action. Dreyfus says, "All attempts to solve [these problems] have run into unexpected difficulties, and this in turn suggests that there may well be in-principle limitations on representationalism" (Dreyfus 1992, p. xviii).

Despite the broad term "representationalism," I think Dreyfus is primarily concerned with the failings of explicit representations in particular. This is suggested by his focus on propositional content in his discussion of representations. But, if this

is right, then tacit representations are not necessarily implicated by his arguments. Dreyfus objects to the "classical" version of AI on the grounds that commonsense and skillful interaction are necessary components of intelligence. But, he argues, commonsense and skillful interaction cannot simply be enumerated, represented, and effectively used by a system.

However, this same enumeration strategy is not required for a system to tacitly represent something. So the insufficiency of enumeration and explicit representation does not necessarily undermine tacit representations in the same way.

I think that adopting a framework that includes tacit representations, rather than just explicit representations, suggests a more promising strategy for meeting Dreyfus's commonsense challenge. Though, of course, many of the details still need to be worked out. But this just means that additional research is required, not that these commonsense problems are insoluble within a framework that includes representations.

25.5.3 A Hidden Lesson

Despite the oft professed repudiation of representation from EC proponents, I think it is more fruitful to understand the lessons for cognitive science and artificial intelligence in a more nuanced way. Our takeaway from EC proponents' discussion of representation should be that *explicit representations* are not up to the task of accounting for some of the foundational components of cognition or intelligence.

The focus of AI work on representation has generally been at a relatively high level of abstraction—e.g. building internal models of the world, using explicit rules, and symbolic logical reasoning. This mirrors the theoretical framework and focus of research in cognitive science. I think we should interpret results from work on embodied cognition as suggesting that cognitive science and AI should shift more focus onto understanding the value of tacit representations for cognition.

One negative project in embodied cognition points out that the classical view, with its focus on explicit representations, cannot account for important cognitive phenomena. But there is a positive project as well. This positive project suggest that important components of cognition are more mechanistic or disposition-based than previously appreciated. I think this positive project actually suggests adopting a framework that includes tacit representations rather than abandoning representations as a way to explain these sorts of cognitive phenomena.

An embodied perspective on cognition should not be used to admonish AI theorists to abandon representations—this is impractical at best and incoherent at worst. Intelligence requires that a system have a way of obtaining and using information about the environment. Artificial intelligence must be designed, so AI engineers must build mechanisms that allow a system to obtain and use information about the environment. But, at some basic level, this is all that is meant by representation—whatever serves the role of conveying information. Advising AI theorists to abandon representations is incoherent if it amounts to a suggestion that

they should give up trying to build artificial systems that obtain and use information about the environment. It is impractical if it amounts to a suggestion that AI theorists should start from scratch in figuring out how to build a system that can obtain and use information.

Instead, the lesson for AI from an embodied perspective on cognition might be more like this: Given what we know about human and animal intelligence and cognition, an important (though previously underappreciated) component for intelligent action seems to be knowledge that is tacitly represented by the system. Work in embodied cognition suggests that some knowledge must be built into the architecture of a system so that the system can display appropriate dispositions.

What I mean by system architecture should be interpreted rather broadly here—both the calculator's circuitry and Brook's subsumption architecture would qualify. Although couched in different terms, connectionist architectures would also fit under this umbrella. Thus, I think one research area embodied cognition suggests is to further investigate the landscape of system architecture within a framework that includes tacit representations. What design principles should guide the design of system architectures such that they are endowed with those tacit representations required for intelligence? How can we integrate system architecture into cognitive tasks?

Another research area that seems especially interesting is how to design systems that can usefully modify their architecture. This ability seems to account for much of the intelligence we find in the animal world. Certainly evolution has resulted in physical architectures that convey important information, allowing organisms to act intelligently. And in the case of humans, our intelligence is often attributed to our comparatively abundant neural plasticity. It would be interesting to consider how to develop analogous abilities for artificial systems.

Continuing in the evolutionary direction, and taking a cue from Brooks' subsumption architecture, it would also be interesting to investigate the foundations and history of cognition in the natural world. Can we identify particular skills that are foundational for intelligent behavior? Is there any particular order in which these skills should be developed or implemented in a system? Studying the evolution of cognition with the tacit representation framework in mind should provide many ideas helpful to developing artificial intelligence that can mimic natural intelligence.

References

Adriaans, P. (2012). Information. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2012.). Retrieved from http://plato.stanford.edu/archives/win2012/entries/information/

Aydede, M. (2010). The language of thought hypothesis. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2010). Retrieved from http://plato.stanford.edu/archives/fall2010/entries/language-thought/

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1), 139–159.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language, 16*(2), 121–145.

Dennett, D. C. (1982). Styles of mental representation. *Proceedings of the Aristotelian Society, 83*, 213–226. doi:10.2307/4545000.

Dretske, F. I. (1983). The epistemology of belief. *Synthese, 55*(1), 3–19. doi:10.2307/20115855.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press.

Dreyfus, H. L., & Dreyfus, S. E. (1988). Making a mind versus modeling the brain: Artificial intelligence back at a branchpoint. *Daedalus, 117*, 15–43.

Floridi, L. (2013). Semantic conceptions of information. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2013.). Retrieved from http://plato.stanford.edu/archives/spr2013/entries/information-semantic/

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1–2), 3–71. doi:10.1016/0010-0277(88)90031-5.

Herford, O. (1899). Retrieved from http://commons.wikimedia.org/wiki/File:Kitty_Riding_Bicycle.jpg

Prinz, J. J., & Barsalou, L. W. (2000). Steering a course for embodied representation. *Cognitive dynamics: Conceptual change in humans and machines*, 51–77. Cambridge, MA: MIT Press.

Shapiro, L. (2010). *Embodied cognition*. Routledge: Taylor & Francis.

Chapter 26
Machine Art or Machine Artists?: Dennett, Danto, and the Expressive Stance

Adam Linson

Abstract As art produced by autonomous machines becomes increasingly common, and as such machines grow increasingly sophisticated, we risk a confusion between art produced by a person but mediated by a machine, and art produced by what might be legitimately considered a machine artist. This distinction will be examined here. In particular, my argument seeks to close a gap between, on one hand, a philosophically grounded theory of art and, on the other hand, theories concerned with behavior, intentionality, expression, and creativity in natural and artificial agents. This latter set of theories in some cases addresses creative behavior in relation to visual art, music, and literature, in the frequently overlapping contexts of philosophy of mind, artificial intelligence, and cognitive science. However, research in these areas does not typically address problems in the philosophy of art as a central line of inquiry. Similarly, the philosophy of art does not typically address issues pertaining to artificial agents.

Keywords Art • Autonomous machines • Philosophy of art • Intentionality • Daniel Dennett • Expressive stance

26.1 Introduction

This paper is framed in relation to Daniel C. Dennett's theory of intentionality and Arthur C. Danto's ontological theory of art. The general structure of my argument is as follows: If, through observation and interaction, we discover an (e.g., artificial) agent's behaviors (and behavioral outcomes) to be similar to those of human agents, this suggests that we may reasonably make similar intentional attributions across agents, understood as intentional systems (Dennett 1987). From this, it follows that, given identical respective outcomes of the behavior of two independent agents (e.g., the production of an artwork by an agent, and of an identical artwork by another agent), we could in principle attribute identical intentions to each agent. However,

A. Linson (✉)
University of Edinburgh, Edinburgh, UK
e-mail: adam.linson@ed.ac.uk

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_26

this equivalent attribution would be problematic in the case of art because two materially identical objects, although indistinguishable in ordinary circumstances, may differ from one another in an important ontological sense, on the basis of their causal origin (Danto 1981). I will argue that for an object to be an artwork, its production must stand in relation to a conscience, which in turn must be based in consciousness. The possibility of consciousness and, ultimately, conscience, is constrained by the design of a cognitive apparatus.

Through a critical examination of Dennett's (1987) idea of the *intentional stance*, applied to the consideration of works of art, I have previously introduced the idea of the *expressive stance* (Linson 2013). According to my account, the intentional stance falls short of an adequate means for understanding art (in relation to artists), and I proposed the expressive stance to address this shortcoming. In the following discussion, I have two primary aims: (1) to further my earlier arguments for the expressive stance with a turn to ontological issues in the philosophy of art, especially those brought forth in Danto (1981); and (2) to tie the ontological issues raised in (1) to a contemporary empirically grounded cognitive neuroscience perspective, in particular, concerning the structure and limits of our broader engagement with the world.

In this paper, I will draw upon ontological issues raised in Danto (1981) to further my arguments for the expressive stance. In doing so, I will draw further contrasts between the expressive and intentional stances, especially with respect to a discussion of machine art. At various turns in this discussion, I will refer to some of Dennett's thought experiments from several of his works to illustrate key points in my argument. Section 26.2, which follows, provides further background to the subsequent discussion. Section 26.3 considers the grounds for different interpretations of materially identical artworks on the basis of their origins, which relate to a sociohistorical context, but also to a cognitive architecture. Section 26.4 considers the relation between an agent's cognitive architecture and the way in which the agent's behavior is interpreted. Section 26.5 examines the connections between conscience, responsibility, and art.

26.2 Background

In my initial sketch of the expressive stance (Linson 2013), I argued that an important aspect of the intentional stance (Dennett 1987) could be preserved when considering works of art (and artistic performances), namely, the epistemological constraint that, in our pragmatic engagement with the world, we are confined to an external view of other subjects; we make judgments through an interpretation of our observations and interactions. However, I also argued that for interpreting works of art in relation to artists, the intentional stance's 'lens' of rationality is insufficient to give a full account of how we understand art as such. My position is that we interpret artworks as an expression of the artist, based on a plausible interpretation of the artist's life experience in a sociohistorical context, grounded in

externally discoverable evidence (in this paper, I will add cognitive architecture to the pool of relevant evidence). This position stands in contrast to the commonly held view that an artwork expresses an artist's intentions, which, in philosophical terms, implies either that the artwork should be understood as resulting from an artist's purported intrinsic intentionality, or, as Dennett (1987) would have it, in terms of a rationalization of the artist's activity, relative (at least partly) to a domain (see Dennett 2001, p. 319ff.). While I find this latter view tenable with respect to general intentional behavior, it cannot on its own account for the ontological aspects of an artwork identified in Danto (1981).

In an ordinary pragmatic context, outside of philosophical discussion, the theorized ontological status of an object, action, or utterance seems to lack practical significance. This is why an externalist account (such as the intentional stance) is useful for explaining the observation of and interaction with a 'black box', from which we can develop a competence theory that need not be a narrow (i.e., Skinnerian) behaviorism (see Dennett 1987, p. 74). Following the logic of a pragmatic, externalist account, I initially formulated the expressive stance without explicitly addressing the ontology of art (Linson 2013). Nevertheless, the context of art brings certain ontological issues into sharp focus.

Danto (1981) investigates the ontological difference between materially identical artworks – or between a materially identical artwork and non-artwork – which he relates to their causal origins and interpretive context. For him, the philosophy of art has as a principal concern precisely these ontological investigations. To set up his discussion, Danto (1981, pp. 1–3) imagines a series of apparently identical red squares from different sources, assembled for an art exhibition:

> The catalogue for it, which is in full color, would be monotonous, since everything illustrated looks the same as everything else, even though the reproductions are of paintings that belong to such diverse genres as historical painting, psychological portraiture, landscape, geometrical abstraction, religious art, and still life. It also contains pictures of something from the workshop of Giorgione, as well as of something that is a mere thing, with no pretense whatsoever to the exalted status of art. (Danto 1981, p. 2)

Danto uses this thought experiment to critically examine a number of perspectives on art, concluding that the definition of art must be a philosophical matter, and more specifically, an ontological one.

Although my previous approach takes a different perspective, I presented an example that is essentially similar to Danto's red squares in my sketch of the expressive stance (Linson 2013). My example concerns an improvising musician's choice between two equally valid notes, which I also compare with the case of two improvising musicians playing what appears to be the same note in otherwise identical conditions. I argue that in either case, a note choice has a role beyond its formal or more broadly rational role (e.g., as the final note of a melody, or as a note placed in the service of crafting a memorable tune); the decision to use a given note connects to the context of the performer's life, of which it is ultimately an expression. While this lens of expression stands in contrast to the rational lens

of the intentional stance, I nevertheless agree with Dennett's (1987) premise in my view that the interpretation of an artwork as an expression in this sense need not appeal to a notion of intrinsic intentionality.

While Danto (1981, Chap. 7) also holds the artist's expression as part of the basis for understanding the nature of an artwork, Danto and Dennett disagree about what is "inside the head" (for instance, of an artist). Danto, a sententialist, finds that Dennett's notion of the intentional stance poses a woefully inadequate challenge to sententialism (Danto 1988). Despite this disagreement, the expressive stance finds support in both Dennett's theory of intentionality and in Danto's theory of the ontology of art. Or perhaps it is more accurate to say that the expressive stance is a critical reconfiguration of both theories: it borrows from Dennett's (1987) view of intentional systems, but highlights a limitation of Dennett's view with respect to art; this limitation is underscored using Danto's (1981) criteria for an ontological distinction among artworks. But, significantly, my argument also indicates that Danto's theory of art does not depend upon sententialism, even though these are intertwined in his own elaboration.

26.3 Empirically Grounded Interpretation

For any artwork, it is uncontroversial to note that some human or nonhuman (e.g., a machine) indeed *physically* produced the work, or at least, through a physical act, offered up a product or process as an artwork (e.g., by submitting it to an art exhibition, holding a public performance, etc.). The expressive stance (following the intentional stance) suggests (*pace* Searle) that we need not posit an originary metaphysically irreducible intentional source for the work, but rather, that our external perspective is all that is needed to form an interpretation of the work. This entails that, given evidence about the physical origins of the work, we can draw an ontological distinction between two materially identical works with different origins, as Danto (1981) points out with his red squares.

To further illustrate the implications of materially identical works with different origins, I will repurpose one of Dennett's typically enjoyable thought experiments featuring Bach and Rudolph the Red-Nosed Reindeer (Dennett 1991, pp. 387–388). As Dennett's version goes, if a previously undiscovered work by Bach were found to have an opening sequence identical to the opening of the Rudolph tune, a present-day listener would not be able to hear the work as an eighteenth century Leipziger would hear it. This is in part because of the obvious associations that we would inevitably make but that they would not, having never heard the Rudolph tune. I would like to use Dennett's example to point out that, in such a case, Bach could not be said to have been *making a reference* to the Rudolph tune, in the everyday sense of a musical reference, as that tune had not yet been written (we will set aside here a dedicated philosophical discussion of reference, which could no doubt lead us in any number of directions). Bach likely would have had a musical justification for composing his melody. Perhaps Bach would have had a similar justification as the

tune's later composer, given their shared human cognitive architecture and relatively similar social experiences. Let us introduce now, to this scenario, a crude robot that merely randomly generates melodies and lacks any additional cognitive apparatus. If the random-melody generator were to produce the Rudolph tune, we could also say without question that the robot was *not* making a reference to the tune. That is, we know Bach could not be making a reference to the tune due to historical evidence, but we know that our robot could not be making a reference to the tune due to evidence about its cognitive architecture.

More importantly, the melody "by" the robot, though apparently identical to that of the unknown Bach piece and that of the Rudolph tune, is ontologically distinct as a piece of music. Although Danto does not take up the issue of cognitive architecture per se, it seems he would agree. Consider his point that a work by Bach differs from a work by a "fugue-writing machine, something that ground fugues out like sausages [. . .]. The person who used it would stand in a very different relationship to the generated fugues from Bach's" (Danto 1981, p. 203). This point acknowledges that an artwork produced by a cognitive architecture similar to our own must be understood differently from an identical work produced by a radically limited collection of mechanisms.

Sloman (1988) makes a related point about cognitive apparatus in response to Dennett (1988; a précis of Dennett 1987). As Sloman states, Dennett "wants the intentional stance to focus entirely on rational behavior and how to predict it, without regard to how the agent is designed, whether by evolution or engineers" (p. 529). One way of understanding this point is to note that, given a certain performance (i.e., behavior or behavioral outcome), we could define a design space of possible models of cognition that would reasonably produce such a performance. The fact that we could go to great lengths to develop an entirely unrelated design that produces an apparently identical performance does not make the unrelated design applicable to understanding the original performance. From this, it follows that we could differentiate between intentional systems in such a way as to not make the same intentional attribution to (e.g.) a human conversationalist as we would to a conversing robot that uses only a look-up table, even if the dialogue were identical in each case.

To further illustrate the idea of an ontological distinction between materially identical artworks, Danto (1981, Chap. 2) turns to Borges' 1939 story, "Pierre Menard, Author of the *Quixote*" (Borges 1998, pp. 88–95), which I will briefly summarize here. In this story, Borges' first-person narrator is a fictional literary critic who reviews the work of a fictional author named Pierre Menard. The world of the narrator and Menard – as in the famous example of Sherlock Holmes' London – intersects with various "non-fictional" aspects of our world, such as the existence of various authors and works, the most central being Cervantes and the original *Don Quixote*. As the narrator tells us,

> Menard did not want to compose *another* Quixote, which surely is easy enough – he wanted to compose *the* Quixote. Nor, surely, need one be obliged to note that his goal was never a mechanical transcription of the original; he had no intention of *copying* it. His admirable

> intention was to produce a number of pages which coincided – word for word and line for line – with those of Miguel de Cervantes. (Borges 1998, p. 91)

Menard thought of two ways of achieving this delightfully absurd goal: "Initially, Menard's method was to be relatively simple: Learn Spanish, return to Catholicism, fight against the Moor or Turk, forget the history of Europe from 1602 to 1918 – *be* Miguel de Cervantes" (p. 91).[1] The narrator tells us that, upon reflection, Menard reconsidered his proposed method, having concluded that "of all the impossible ways of bringing it about, this was the least interesting. [. . .] Being, somehow, Cervantes, and arriving thereby at the Quixote – that looked to Menard less challenging (and therefore less interesting) than continuing to be Pierre Menard and coming to the Quixote *through the experiences of Pierre Menard*" (p. 91, original emphasis).

According to the narrator,

> it is a revelation to compare the *Don Quixote* of Pierre Menard with that of Miguel de Cervantes. Cervantes, for example, wrote the following (Part I, Chapter IX):
>
> . . . truth, whose mother is history, rival of time, depository of deeds, witness of the past, exemplar and adviser to the present, and the future's counselor.
>
> [. . .] Menard, on the other hand, writes:
>
> . . . truth, whose mother is history, rival of time, depository of deeds, witness of the past, exemplar and adviser to the present, and the future's counselor. (p. 94)

Comparing these excerpts, the narrator describes the "striking" contrast in styles: "The Cervantes text and the Menard text are verbally identical, but the second is almost infinitely richer" (p. 94).

Danto (1981) uses this story to underscore his point about the materially identical red squares introduced earlier, noting that the conditions under which they were produced – their causal origin – is situated in a sociohistorical context, just as the texts by Cervantes and Menard. His point is that our *interpretation* of artworks must take their sociohistorically situated causal origins into account:

> You can certainly have objects – material counterparts – at any time in which it was technically possible for them to have come into existence; but the works, connected with the material counterparts [. . .], are referentially so interlocked into their own system of artworks and real things that it is almost impossible to think of what might be the response to the same object inserted in another time and place. (Danto 1981, p. 112)

And while Danto (1981) does not explicitly consider cognitive architecture, I have indicated how this consideration could be relevant to understanding the difference between two works produced under similar sociohistorical circumstances by agents with vastly different cognitive architectures.

---

[1] Dennett makes a similar proposal (which he dismisses as unnecessary) for how one might try to experience a Bach cantata as an eighteenth century Leipziger would have experienced it: "To put ourselves into the very sequence of experiential states such a person would enjoy [. . .] would require [. . .] forgetting much of what we know, losing associations and habits, acquiring new habits and associations". This would take place in "isolation from our contemporary culture – no listening to the radio, no reading about post-Bach political and social developments, and so forth" (Dennett 1991, pp. 441–442).

26.4 Conduct and Context

The relation between behavior and interpretive context will be explored in this section in relation to cognitive architecture, personhood, and computational creativity. As indicated in the previous section, we may consider the relation of cognitive architecture to the interpretation of an agent's behavior. Cognitive architecture, though technically "inside the head" by some accounts, does not necessarily point to intrinsic intentional states. Significantly, an agent's cognitive architecture may be relevant to the determination that the agent is a person, even without an appeal to mental content (cf. Dennett 1981, Chap. 14). An agent may have specific capacities that allow it to engage in humanlike behavior, while nevertheless lacking the capacities that underpin personhood. When artificial agents are designed to exhibit creative behavior, as in some computational creativity research, a specific aim may be to produce artworks. The philosophical dimension of these artworks will be considered below.

As a way into this discussion, I will use another of Dennett's thought experiments, which I will briefly summarize here. Dennett also reflects on a story by Borges – in this case, "The Circular Ruins" – to set up his premise: Suppose there is a "novel-writing machine, a *mere* machine, without a shred of consciousness or selfhood" (Hofstadter and Dennett 1981, p. 351). Dennett adds that we can suppose its designers "had no idea what novels it would eventually write". He develops the thought experiment further, coming to the point at which, rather than a simple novel-writing box-like machine, we are faced with a (presumably humanoid) robot that speaks aloud a first-person narrative. Its spoken narrative more or less corresponds to what we can observe about it (e.g., "When it is locked in a closet, it says: '*I* am locked in the closet!'", p. 351). With this setup, Dennett then poses the question: Why should we call this spoken narrative fictional? He answers that we should not, given that this is effectively how human brains work: "Your brain, like the unconscious novel-writing machine, cranks along, doing its physical tasks [. . .] without a glimmer of what it is up to. [. . .] *It* doesn't 'know' it is creating you in the process, but there you are, emerging from its frantic activity" (pp. 351–352, original emphasis). If we go along with this, Dennett thinks, we should also be willing to conclude that the robot is a *person*, on the basis that its personhood emerges from its "activity and self-presentation in the world" (p. 351).

There is something very important about interpretation that is missing here, which is surprising, given Dennett's (1987) insightful foregrounding of the role of interpretation in his theory of intentionality. While we can indeed take into account an agent's "activity and self-presentation in the world" for making judgments about personhood, the context for interpreting the agent's behavior is crucial to making the distinction between a fictional entity and a person. Danto (1981, p. 23) recognizes this distinction when he identifies the interpreted difference between an assertion and a *mention of an assertion*, i.e., an assertion in quotation marks; he links this with our ability to identify the conventional context of a theatrical play, for example,

which guides our interpretation of actions and words on a stage. Such interpretations must differ from those of actions and words outside of the theater.

We may reasonably go along with the idea that Dennett's robot is a person in one of two ways, both of which seem to undermine the point he seeks to make with his scenario. In the first way, we may take a vantage point from which we have no knowledge that we are dealing with a robot, thus mistaking it for a human and bringing all of our assumptions about humans to bear on the situation. In this case, its activity and self-presentation are already interpreted as relating to personhood, even if this interpretation would be importantly altered with key facts about the robot's cognitive shortcomings. Or, in the second way, we may have foreknowledge that we are dealing with, say, a highly sophisticated robot, wired up like a human and perhaps even having a human-like upbringing, such that by some standards of human equivalence, it already may be considered a person, regardless of its present activity and self-presentation, e.g., if the robot is somehow temporarily deactivated.

This point can be made more clearly using a different example. Imagine a performance artist sitting completely still in a room. With no knowledge about the performer or context, the performer's mere activity and self-presentation would not necessarily lead anyone to believe they are a person. The determination that they are a person could be made either with some brain scans and other medical tests that would satisfactorily prove they are a living human; this would at least imply that they should probably be considered a person ("we normally [. . .] treat humanity as the deciding mark of personhood", although they may be exceptions; see Dennett 1981, p. 267). Or, we could be furnished with the context that we are witnessing a performance in a museum by someone operating within a certain performance tradition, and so forth, in which case we could assume that their motionless sitting *is* in fact their activity and self-presentation in the world. This interpretation requires the broader context, which would reasonably lead us to conclude they are a person. In neither case would a robot known to have crude cognitive machinery, designed for superficial humanlike activity and self-presentation, be considered a person.

While I am in agreement with Dennett (1981) that the conditions of personhood should not be grounded in intentional states, I would, however, respond that its conditions may nevertheless be grounded in capacities such as having experience, making judgments, reflecting, and so on, which certainly some future machine may be capable of, but hardly an unconscious novel-writing machine. Empirical questions about these capacities can be explored with respect to the cognitive architecture that underpins them (see Sloman 1988). Significantly, these capacities relate to our broader engagement with the world, while leaving in tact the idea that they may complement domain-specific cognitive mechanisms for particular specialized activities (e.g., writing a novel).

We must, however, tread carefully in our understanding of the relation between these general capacities and more specific ones, such as creativity. One of the key ideas behind computational creativity is that there are cognitive mechanisms associated with the production of new ideas and, by extension, with the production

of new works of art. Boden (1990/2004) distinguishes between what she terms personal or psychological creativity – meaning a new idea arises in an individual cognitive agent – and historical creativity, meaning that an idea is recognized as new and valuable by a community. Research in this area often focuses on psychological creativity though the development of machines with a high degree of autonomy that produce art, especially musical, visual, or literary art. In this context, the distinction can be easily obscured between some of the domain-specific activity relevant to artistic production and the general cognitive activity relevant to personhood. For example, a short story writing machine developed by Bringsjord and Ferrucci (1999, p. 100) has the stated aim that it "holds its own against human authors" (in terms of observable behavior). Their machine's approach "to story generation is based on [overcoming] the assumed limitation of computers to genuinely grasp such things as interestingness" (p. 199). In this case, the ability to identify an interesting story is regarded as a domain-specific capacity rather than a general one.

Given that the idea of an unconscious novel-writing machine is clearly not as far-fetched as it might sound to some, in this context, I would like to consider a distinction between a work produced by an expert system and a work produced by a person. Following points by Danto (1981) and Sloman (1988) introduced above, we may say that materially identical works, one produced by an expert system and the other by a person, must have a different ontological status as artworks, because of the way the work relates to the conditions under which it was produced. These conditions must be taken into account for a defensible interpretation of the work. As Danto states,

> It is not just that appreciation [of an artwork] is a function of the cognitive location of the aesthete, but that the aesthetic qualities of the work are a function of their own historical identity, so that one may have to revise utterly one's assessment of a work in the light of what one comes to know about it; it may not even be the work one thought it was in the light of wrong historical information. (Danto 1981, p. 111)

This relates to the present discussion in that, if we are moved by (e.g.) a novel and think it captures something about human experience, but we then discover that it was produced by a robot with a cognitive apparatus that is limited in certain key respects, this new information demands of us that we modify our original interpretation. We may still agree that the story moved us, but not because the robot shares an idea about human experience.

In fact, in this example, it is the robot's designers who share with their audience an idea about human experience. They may have designed, for instance, a largely autonomous system that generates stories and decides that they are complete and ready for the public. In doing so, however, the designers have made a decision of their own that significantly impacts what the robot does. Namely, they have decided that their purpose-built robot is itself complete and ready for the public. The designers' judgment that their art-producing robot is ready to be 'released' is part of their judgment that its output objects can, from that point forward, be treated as artworks (novels, paintings, performances, etc.). Even if such a system has an internal 'critic' to evaluate its own output, the criteria for the critic have likewise been established by the designers (though computational implementations

of aesthetic evaluation may be "almost comically faulty"; see Thomson 2011, p. 60). The designers thus bear aesthetic responsibility for the machine output, even if they do not know the precise objects it will produce.

My position is that any such work, despite the designers' lack of knowledge about future system output, is nevertheless an expression of the designers, rather than of the machine. If the works are regarded as artworks, they must be understood as works by human artists, mediated by autonomous machine production. This is importantly different from an actual machine artist, which would only be possible if the machine were a person (as some future machine may be, but, as far as we know, no current machine is). Thus, mine is not among the familiar positions that art could only be art if made by a human, or that a machine could never be creative, have emotions, etc., which often seem to be the main positions defended against in this context by philosophers including Dennett and Boden.

Returning to Dennett's robot, let us assume it is highly sophisticated, beyond all current technology, and that we could have a conversation with it that is apparently about its childhood. Even if it could pass a Turing test for intelligence, there remain two important facts for us to know about the robot before we ought to be tempted to consider it a person. The first fact concerns its location in history: Was it powered up today for the first time, revealing its designer's remarkable generator of childhood stories? Or, did it actually spend time as a less-developed robot with a humanlike upbringing, analogous to a human child? While this may seem an obvious piece of discoverable evidence, it remains outside of the Turing test, as the machine could always give answers as if the latter were the case, even if the former were true. The second fact concerns its cognitive architecture: Does it use a cognitive architecture similar to our own, capable of experience, judgment, reflection, etc., or does it merely use a crude look-up table or similarly simplistic mechanism? As Sloman (1988, p. 530) points out, there are "*design* requirements for various kinds of intentional abilities" and certain philosophical considerations are mistakenly based on "an oversimple view of the space of possible designs" (see also Shieber 2014).

While an external view of a 'black box' may suffice for understanding a limited practical engagement (typical of our daily encounters with others), the answers to the above questions would be evidence that gives us more or less justification to treat the robot as significantly equivalent to ourselves – perhaps even as a person – regardless of its inherent biological difference. Assuming we do find it to have autonomously developed over time, as part of our society, accruing experience, exercising a faculty of judgment, undergoing reflection, etc., if it *then*, under these circumstances, were to produce an artwork (write a novel, paint a painting, improvise a musical solo, etc.), we would be in a position to reasonably interpret its work as that of a machine artist, despite some prior role of a robot designer. Such a robot would not be regarded as a traditional domain-specific expert system, but more like what we would ordinarily regard as a person. Its artistic output would be ontologically different than randomly-generated output or output generated because its designers settled upon domain-specific mechanisms to produce even largely unforeseeable works.

26.5 Conscience and Consciousness

As I have argued, artificially intelligent art-producing machines can be regarded as a class of expert systems, but works of art as such (as opposed to objects identical to artworks) cannot be adequately understood as mere exercises of specialized skills. Rather, such skills, even when local to the production of a specific artwork or performance, fundamentally relate to a broader engagement with the world. Thus, the critical difference between a domain-specific expert system and a person, as viewed from the expressive stance, can in some sense be understood as a reassertion of Dreyfus' (in)famous Heideggerian critique of artificial reason (Dreyfus 1992; see also Dreyfus 2008), but narrowly focused on art.

With respect to art and intentionality, in Linson (2013) I explicitly draw on Dreyfus' (1993) Heideggerian critique of Searle's two senses of the phenomenological "Background". In short, the critique holds that our actions are not only relative to our physical bodies and cultural circumstances, but also to our broader engagement with the world. Roland Barthes makes a similar point when he argues that it is not only one's body and sociohistorical circumstances that give an author (or artist) a unique perspective, but also a decision to have a particular take on the world (Barthes 1968). For Barthes, this take is linked to the idea of conscience.[2] Dennett (1981, p. 297) also accords an important role for something like conscience when he partly locates our moral responsibility in the fact that we decide when to terminate our deliberation process for executing a given action.

Conscience is not an especially well-defined term in philosophy, and some view it as an inner voice that distinguishes between what is morally right and wrong. But this view is too narrow for what I am trying to capture here. I am instead suggesting that conscience is relevant even when not faced with a straightforwardly moral question. In some sense, any social act has an inherently ethical dimension, and our decisions to act in certain ways relate to our general sensibilities about how we ought to act; these sensibilities are partly arrived at through socialisation and partly arrived at through self-examination and reflection. Making art of any kind (not only overtly political art) is not the type of activity that is usually held up as a moral act, but it may be viewed as the outcome of an artist's deep convictions about humanity. As philosopher of law Larry May (1983, p. 66) states, conscience motivates us "not to view our own selves as the end to be served, but to view humanity *in* our own and other persons as the end to be served" (original emphasis). An artist's convictions of this sort pertain not only to the production (or performance) of a specific work, but also pertain to the more general decision to produce art for the public, in other words, the decision to be an artist.[3]

[2]It is interesting to note that the French word, *conscience*, may be used to mean either conscience or consciousness.

[3]Conscience also plays a role in Danto's (1981) contention that artists must be morally responsible in their decisions about what to portray and how to portray it, which he explores in relation to the concept of the "psychic distance" an aesthetic attitude has from a practical one (pp. 21–24).

Before addressing conscience further, we may note that Dennett and Barthes agree that one's consciousness (in the general sense of subjectivity) is affected by the society and historical period in which it developed. To take one example, Barthes (1968) points out a difference between the writing styles of Balzac and Flaubert that he attributes to the fact that their lives are separated by the events of revolutionary Paris in 1848. The societies and traditions of these literary figures are, on one hand, remarkably similar, but, on the other hand, their respective worlds differ significantly. It seems Dennett would agree, on the basis of a related example:

> There are probably no significant biological differences between us today and German Lutherans of the eighteenth century [...] But, because of the tremendous influence of culture [...] our psychological world is quite different from theirs, in ways that would have a noticeable impact on our respective experiences when hearing a Bach cantata for the first time. (Dennett 1991, p. 387)

As Dennett suggests, when we hear Bach's chorales today rather than in Bach's time, "we hear them with different ears. If we want to imagine what it was like to be a Leipzig Bach-hearer, it is not enough for us to hear the same tones on the same instruments in the same order; we must also prepare ourselves somehow to respond to those tones with the same heartaches, thrills, and waves of nostalgia" (Dennett 1991, p. 387).

Generally speaking, whether we are concerned with short stories or novels, musical compositions or improvisations, performance art or any other artistic media, we regard artworks as a result of decisions by one or more artists – at the very least, deciding when a work is finished and ready for the public, though we may also include decisions about formal and structural aspects of the work and, at a more "global" level, decisions such as the adherence to or flouting of traditions, etc. From the vantage point of the expressive stance, we can read such works as expressing something – whether their author intended them to or not – about the times and society in which the author lived, and the experiences the author underwent in these contexts. In this sense, we may say the works are an expression of the artist's life and, in particular, of the artist's consciousness. But if the decision-making apparatus of an individual artist is relevant to the interpretation of an artwork, how can we adopt an external vantage point that disregards the traditional notion of authorial intentions "inside the head"? And, assuming we disregard such intentions, how might this relate to the theoretical notion of the "death of the author"?[4]

It was in fact Barthes himself who, in a 1967 essay, "The Death of the Author", introduced its titular notion in the contemporary sense (Barthes 1977). Barthes used the phrase in response to an earlier generation of literary critics who believed that there must be one ultimate meaning of a given literary work (to which we may also add, any artwork). This meaning was assumed to be deducible from the author's intentions, conscious or unconscious, public or hidden, which could, for example, be partly uncovered in the author's journals, biography, and so on (a version of this

[4]This concept was mentioned briefly in Linson (2013) but, due to space limitations, was not addressed in depth.

view is known by literary critics as the "intentional fallacy"; see also Dennett 1987, p. 319). Barthes argued that a multiplicity of interpretations should be possible, because the meaning is partly constituted by the reader, who brings his or her thoughts and experiences to bear on the interpretation of the material. We can make sense of this view in relation to other artistic practices as well: "responding to a painting complements the making of one, and spectator stands to artist as reader to writer in a kind of spontaneous collaboration" (Danto 1981, p. 119).

The idea of the death of the author – much like Dennett's (1987) critique of intrinsic intentionality – is to delink the (outwardly observable) work from some imagined definitive "theological" key in the author's mind that holds the ultimate answer to the intentions behind the work. According to Barthes' view – which is also similar to Dennett's (1981, 1987, 1991) critical view of introspection – the author's intentions, which were held by prior critics as constituting the originary source of an artwork, should not be assumed to be consciously furnishable by the author, say, in an interview, nor should they be assumed to be discoverable by the critic researching the author's memoirs. Rather, we should not understand any definitive intentions as being ultimately expressed in or by the work. Coming to an understanding of an artwork – or hermeneutically engaging with its meaning, without a notion of a final point of arrival – is a process of interpretation that is grounded by evidence. There may be irresolvable conflicts among competing interpretations, but this situation does not fundamentally differ from that of the sciences: As Danto (1981, p. 113) notes, there is a "slogan in the philosophy of science that there are no observations without theories; so in the philosophy of art there is no appreciation without interpretation".

The mode of interpretation proposed by Barthes (1977) is very close to Dennett's position on intentional interpretation, where we may encounter a number of plausible reasons for someone's decision leading to an action, reasons which can never be definitively proven, although a better or worse case can be made (see Dennett 1987, Chap. 4). For Dennett, this scenario exemplifies taking the intentional stance: What is "inside the head" is not pragmatically relevant to the interpretation of an intentional system's activity or output. Rather, we interpret such activity and output on the basis of a community of shared meaning. Thus, Barthes' "death of the author" is already to some extent a preliminary version of the intentional stance epistemology for the arts. But, importantly, this interpretation-centric view is not a pejoratively construed "anything goes" version of postmodern theory. A reasonable basis for interpretation is important, just as Dennett (1987, p. 100) points out for interpreting Jones' likely delusional rationale for the terrible-looking extension to his house. As with scientific observation, there may be facts that undermine certain interpretations (recall Danto's (1981) point that "one may have to revise utterly one's assessment of a work in the light of what one comes to know about it", p. 111).

We have already indicated the importance of evidence about an artwork's origins for determining its ontological status. We have also pointed out that those origins at once relate to sociohistorical location, cognitive architecture, and, at the intersection of both, the artist's consciousness – and, more specifically, the artist's conscience, an important aspect of personhood. A person – natural or artificial – can be said to

produce an artwork through a process of making decisions, not only about the form and content of the work, but about when it is ultimately ready for the public. Dennett (1981, p. 297) makes a related point concerning the basis of our responsibility for moral decisions: "In many cases our ultimate decision as to which way to act" is importantly connected to "prior decisions affecting our deliberation process itself: the decision, for instance, not to consider any further, to terminate deliberation; or the decision to ignore certain lines of inquiry". He continues:

> These prior and subsidiary decisions contribute [. . .] to our sense of ourselves as responsible free agents, roughly in the following way: I am faced with an important decision to make, and after a certain amount of deliberation, I say to myself: "That's enough. I've considered this matter enough and now I'm going to act," in the full knowledge that I could have considered further [. . .] but with the acceptance of responsibility in any case. (Dennett 1981, p. 297)

When an artist finally decides to declare a work as ready, this constitutes the termination of a cycle of on-going action and judgment and the taking of responsibility for the production of the work. Here, it is true that, from the intentional stance, we can understand the work in one way by an appeal to reason: the painter decided to add no more, so the canvas would not be overworked; the poet decided to stop editing the poem, so its initial impulse would not be obscured. But apart from an explicit or attributable rationale, the decision to conclude with the production of a work is shaped by the artist's sensibility, a sensibility that arises from the life experience of the artist and, ultimately, from the artist's conscience, formed in the course of experience. This is not to deny that the life experience of artists is in many ways due to factors beyond their control, including sociohistorical circumstances of geography and culture, contingencies about their physical bodies, and the evolutionary biological inheritance that amounts to a cognitive apparatus. Within these constraints, however, a conscience is formed that allows one – an artist or indeed any person – to take responsibility for one's decisions, such as those pertaining to an artwork.

26.6 Conclusion

We are now in a position to account for the question posed in the title: Machine art or machine artists? As I have argued, we may identify how an artwork relates to a specific historical period and culture, a specific artist's body and cognitive apparatus, and, subject to these constraints, an artist's conscience. A conscience in this sense is certainly historically situated and embodied, but also relates to a person's unique accumulation of experience. One develops such a conscience and it affects one's decisions about how to carry out certain activities, such as producing an artwork and determining that it is ready for the public. The development of a conscience is only possible given particular facts about our cognitive apparatus that facilitate our accumulation of experience, our faculty of judgment, our capacity for reflection, and so on. Thus, my argument entails that for a machine-produced

artwork to be regarded as a contribution by a machine artist – rather than by the machine's designers – the machine's cognitive architecture must make possible experience, conscience, and the closely related broad engagement with the world. Without such a cognitive architecture, machine-produced art must instead be understood as the work of a human artist, mediated by a machine.

Danto (1981) has confined his account of an artist's relevant cognitive processes to a sentential account of the mind, which, following my above arguments, should be viewed as entirely unnecessary to his ontology of art. An empirically based neurobiological account of our cognitive architecture, for instance, could plausibly describe how the mind might call upon experience to guide action without appealing to a sentential account. By dispensing with sententialism and taking cognitive architecture into account, the expressive stance can facilitate the ontological distinction between artworks and non-artworks that Danto envisions, while preserving Dennett's insight that intentional states should not be understood as intrinsic, as I have argued with respect to an artist's intentions.

Along the lines envisioned by the Turing test, we may say that, using the best available evidence during strictly external observation and interaction – that is, without an appeal to what is "inside the head" – we may potentially detect no difference between a human and a machine. If an unknown entity is determined to be sufficiently adequate at conversing, playing chess, etc., then we are reasonably entitled to make certain assumptions about it (e.g., that it can think). For example, during an ordinary form of interaction with a neighbor at the local store, we may set aside the question as to whether this neighbor is indeed just like us, or whether they are perhaps a robot (or a zombie, etc.). The neighbor, like us, is an intentional system, that is, a system to which we can consistently attribute intentional behavior.

However, a look inside an intentional system – at its cognitive apparatus, not at some metaphysical notion of mental content – could reveal a relatively crude apparatus, like a look-up table. This discovery may render false our prior assumptions about the agent and thereby lead us to draw different conclusions about its personhood, even if we would ordinarily grant it personhood on the basis of our external interactions and observations alone. Assuming equivalent external performances, an agent with a cognitive apparatus similar to our own should be regarded as more deserving of the ascription of consciousness than a look-up-table-based agent. If consciousness can be reasonably attributed to the agent, we may then inquire into what mechanisms structure the agent's decision-making, and whether or not these can be said to ultimately relate to a conscience that underlies the agent's ability to take responsibility for (e.g.) producing an artwork. To determine whether an artwork is the expression of a machine or its designer, we must recognize the fundamental relationship between an artwork and a conscience. Ontologically speaking, an object can only be an artwork because of this relationship.

References

Barthes, R. (1968). *Writing degree zero*. New York: Hill and Wang.
Barthes, R. (1977). *Image, music, text*. London: Fontana Press.
Boden, M. (2004). *The creative mind*. London: Routledge [1990].
Borges, J. L. (1998). *Collected fictions*. New York: Penguin.
Bringsjord, S., & Ferrucci, D. (1999). *Artificial intelligence and literary creativity*. Hove: Psychology Press.
Danto, A. C. (1981). *The transfiguration of the commonplace*. Cambridge: Harvard University Press.
Danto, A. C. (1988). The notional world of D. C. Dennett. *Behavioral and Brain Sciences, 11*, 509–511.
Dennett, D. C. (1981). *Brainstorms*. Cambridge: MIT Press.
Dennett, D. C. (1987). *The intentional stance*. Cambridge: MIT Press.
Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences, 11*, 495–505.
Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown and Co.
Dennett, D. C. (2001). The evolution of culture. *The Monist, 84*(3), 305–324.
Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge: MIT Press.
Dreyfus, H. L. (1993). Heidegger's critique of the Husserl/Searle account of intentionality. *Social Research, 60*(1), 17–38.
Dreyfus, H. L. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland, & M. Wheeler (Eds.), *The mechanical mind in history* (pp. 331–371). Cambridge: MIT Press.
Hofstadter, D. R., & Dennett, D. C. (1981). *The mind's I*. New York: Basic Books.
Linson, A. (2013). The expressive stance: Intentionality, expression, and machine art. *International Journal of Machine Consciousness, 5*(2), 195–216.
May, L. (1983). On conscience. *American Philosophical Quarterly, 20*(1), 57–67.
Shieber, S. M. (2014). There can be no Turing-test-passing memorizing machines. *Philosophers' Imprint, 14*(16), 1–13.
Sloman, A. (1988). Why philosophers should be designers. *Behavioral and Brain Sciences, 11*, 529–530.
Thomson, I. D. (2011). *Heidegger, art, and postmodernity*. Cambridge: Cambridge University Press.

Chapter 27
Perception, Action and the Notion of Grounding

Alexandros Tillas and Gottfried Vosgerau

Abstract Traditionally, philosophers and cognitive scientists alike considered the mind as divided into input units (perception), central processing (cognition), and output units (action). In turn, they allowed for little – if any – direct interaction between perception and action. In recent years, theorists challenged the classical view of the mind by arguing that bodily states ground cognition. Even though promising, the notion of grounding is largely underspecified. In this paper, we focus on the debate about the relation between perception and action in order to flesh out the process and in turn clarify the notion of grounding. Given that currently the debate about the relation between perception & action is far from settled, we attempt an assessment of the implications that possible outcomes of this debate would have on Grounding Cognition Theories. Interestingly, some of these possible outcomes seem to threaten the overall program of Grounded Cognition. In an attempt to make this analysis more concrete, we study two closely related speculative hypotheses about possible ways in which perception and action interact. Namely, we focus on Theory of Event Coding and Simulation Theory, and evaluate the levels of compatibility between those two views and Grounded Cognition Theories.

Keywords Perception • Action • Embodiment • Grounding • Grounded cognition

27.1 Introduction

Traditionally, philosophers and cognitive scientists alike considered the mind as divided into input units, central processing and output units. In this view, perception and action (henceforth P&A) are playing the input and output roles respectively, while cognition, the central processor unit, is responsible for the remaining processes. Urged by the aforementioned vertically modular structure of the mind, Hurley (2001) examines these classical views about the structure of the mind under

A. Tillas (✉) • G. Vosgerau
University of Düsseldorf, Düsseldorf, Germany
e-mail: atillas@phil.uni-duesseldorf.de; vosgerau@phil.uni-duesseldorf.de

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_27

the umbrella term 'Classical Sandwich Model'. Proponents of this model allow for little – if any – direct interaction between P&A and thus draw little attention to the relation between P&A. Recently, this view of the mind has been questioned and various – often opposing – views about the ways that P&A interact have been put forth. For instance, proponents of Grounded Cognition – a part of the Embodied Cognition view – argue that cognition is grounded in bodily states (cf. Barsalou 2008; Shapiro 2007; Wilson 2002).

Our main target in this paper is to examine the debate about the relation between P&A in order to further flesh out the process of grounding. The motivation for examining the relation between P&A stems from the central claim behind Grounded Cognition Theories (henceforth GCTs), that sensorimotor representations ground cognition.[1] In the interest of simplicity, we start from the intuitive claim that sensory representations are involved in perceptual processes and motoric representations are involved in motor processes. It is worth clarifying at this point that we do not imply a 'vertically modular structure' of the mind in the way that 'classical sandwich' views do. We merely use it as a starting point in order to assess the relation between P&A. At the end of the paper, we will be in a position to shed light on the relation between P&A, as well as on the relation between sensory and motor representations. Furthermore, we will argue that different theoretical outcomes of the P&A relation debate have different implications for GCTs.

27.1.1 Preliminaries

As mentioned above, the notion of grounding is underspecified. In order to focus the discussion, we suggest that the claim that 'bodily states ground cognition', on a plausible and accessible first construal, means that cognition is analyzable in bodily states or that bodily states underlie cognition or cognitive processes.

In examining the ways in which perception and action interact, we consider two candidate theories, 'Theory of Event Coding' (TEC) and 'Simulation Theory' (ST)

[1] In this paper, we proceed within a broadly construed computationalist framework and focus our analysis on sensory and motoric representations. However, we do not argue for or commit to neither strong nor modest computationalism. Clarifying this point is crucial given that grounded and embodied cognition views – especially in their stronger versions – are often seen as alternatives to Computational Theory of Mind. At the same time, even though we are sympathetic to the idea of grounded cognition, we do not commit to such views either. Furthermore, we use the notion of representation in a much broader sense than the one used in classical computational views (i.e. symbol), and we do not commit to the claim about all aspects of cognition being representational. To this extent there might well be aspects of cognition that are non-representational and non-computational (a view that modest computationalists would probably accept – see Horst (2009) for a discussion). Our focus here is simply to assess and evaluate the implications that the developments in the debate about the relationship between perception and action have for grounded cognition theories.

(see Sects. 27.3 and 27.4 below respectively). Having a working hypothesis about how perception and action interact is crucial in evaluating the implications that different outcomes of the P&A relation debate have for the Grounded Cognition research programme. In our reading, TEC and ST are two of the most plausible available views about the ways that P&A interact. However, both of those theories rely on speculative hypotheses about the structure of the mind to which our analysis is not committed.

27.2 How Are P&A Related?

Sensory and motor representations are clearly distinct to the extent that they are involved in distinct processes, perceptual and motor processes respectively. Despite these differences though, sensory and motor representations relate closely and are involved in interacting processes. For instance, it is intuitive to assume a causal interaction between sensory and motor states, in the sense that there are contingent causal interactions or influences between the two. However, and as shown below, it is more difficult to have a clear-cut answer about the ways in which P&A interact, as well as about what specific influences motor representations have on sensory representations and vice versa.[2]

There are two main strands in the debate about the nature of the relation between P&A. On the one hand, there are views according to which action is 'constitutively' involved in perception. Largely, 'constitutive relation' in this literature means that you cannot have one without the other and vice versa. For instance, O'Regan and Noë (2001) argue that in order to perceive a given stimulus, we need to register the 'sensorimotor contingencies' related to this stimulus. More clearly, in order to perceive a given object we need to be in a position to anticipate how the visual stimulus will change were we to move our eyes or bodies (or was the object to be moved). Sensorimotor dispositions underlie this implicit knowledge. Noë (2005), in particular, argues that without these dispositional motoric responses to visual inputs there would be no perceptual content. In a similar fashion, amongst others, Kiverstein (2010) appeals to sensorimotor expectations and argues that the content of perceptual experience is of a sensorimotor nature.

On the other hand, there are views according to which P&A are functionally distinct. Proponents of such views most often appeal to evidence from the two visual streams literature (e.g. Ungerleider and Mishkin 1982; Milner and Goodale 1995, 2004, 2010). More specifically, Jacob and de Vignemont (2010) suggest that despite interacting strongly, visual perception and visuo-motor behavior are functionally distinct. Similar views can be found, amongst others, in Jacob and Jeannerod (2003) who argue that P&A are cognitively distinct. Gallese (2007) argues that the dorsal

[2]See Gangopadhyay et al. (2010) for a detailed discussion about the nature of the relation between P&A.

and the ventral streams (see below) are partially dissociated, and thus implies a more moderate view about the relation between P&A.

At the current state of the debate, there is no conclusive evidence either for a constitution relation (in the sense that you cannot have one without the other) holding between P&A, or for a functional dichotomy holding between the two. The only claim that one could safely put forth is that P&A strongly interact. A plausible way to construe the way that P&A interact is in terms of the role that eye movements play in the visual function. For instance, Findlay and Gilchrist (2003) argue that eye movements are crucial for visual perception. They base their claims on evidence from examining AI, a subject who cannot make eye movements. AI's inability to move her eyes does not have a dramatic impact on her visual perception, but only because AI uses head movements to compensate for the absence of eye movements. The (neural adaptation – probably in the superior colliculus – responsible for the) transfer of saccadic movements from the eyes to the head exhibited in the case of AI, is suggestive of the significance of eye movements for visual perception.

Despite the lack of conclusive evidence concerning the nature of the relation between P&A though, evaluating the theoretically possible outcomes of this debate could shed light on how sensorimotor representations ground concepts (or cognition).[3] In particular, instead of going through different arguments for and/or against a constitution relation or a functional distinction between P&A, we will take into account two main intersecting points and use them in order to evaluate the nature of sensorimotor representations, which allegedly ground cognition. Namely, we focus our analysis on (i) the nature of the relation between P&A and (ii) on the computational level at which P&A interact or converge. As already mentioned, the theoretically possible outcomes of the P&A relation debate are that either a constitution relation – in the sense that you cannot have one without the other – or functional distinction holds between the two. Concerning the computational levels at which this interaction could occur, we distinguish between a higher level involving concepts, and a lower level involving only sensory and motoric representations, and examine two possible ways that P&A could interact. In doing so, we examine Hommel et al.'s 2001 Theory of Event Coding (TEC); and Barsalou's (1999) Simulation Theory (ST). As already mentioned, both of those theories are speculative hypotheses yet remain the most plausible candidates – at least in our reading of the relevant literature. It is also worth clarifying that even though both of these theories build on associationism, they largely differ. In what follows, we treat ST as a more clear-cut case of associationism since it does not treat sensory and motor representations as commonly coded, but as 'merely' associated.

[3]Further elaboration on the relation between 'constitution' and 'grounding' extends beyond the scope of the present paper. See Weber and Vosgerau (2012) for a detailed discussion.

27.3 Theory of Event Coding (TEC)

One of the most prominent views in the literature about the way that P&A interact is Hommel et al.'s (2001) Theory of Event Coding (TEC). TEC uses Prinz's (1997) Common Coding Theory (CCT) as a theoretical background. Briefly, according to CCT, sensory and motoric information is encoded in the same representational format. That is, actions are coded in the mind in terms of the perceivable effects they should generate (we use 'formats' and 'codes' here interchangeably). More generally, a given bit of information is representable in different ways regardless of how it is realized. For instance, according to CCT, the action of lifting a cup is coded in terms of the visual representations of the cup being lifted.

Before elaborating on TEC, it is worth clarifying further that aspects of TEC do enjoy significant empirical support. However, it is not as yet empirically fully vindicated and rather remains a speculative theory about the relation between P&A to which we are not committed. Our target is simply to evaluate the implications for Grounded Cognition theories, should it turn out that P&A interact the way suggested by TEC.

TEC does not cover all kinds of possible interactions between P&A – the focus here is on action planning. In more detail, TEC theorists understand perception as the late *cognitive* products of perceptual processing. Qua being cognitive, these products represent specific features of actual events in the subject's environment, and not the preceding sensory processes that eventually lead to them. At the same time, they understand actions as the early cognitive antecedents of action that represent specific features of actions. In this sense, TEC focuses at a high cognitive level. This is of great significance in assessing the implications of running Theory of Event Coding and Grounded Cognition projects together, as shown in Sect. 27.5.3.

Perception and action planning interact strongly and this interaction often produces indistinguishable processes. However, perception and action planning processes intersect only in cases in which the codes they operate on refer to the same (kind of) features of a given distal event. Planning an action, in this sense, does not involve specification of a detailed muscular activity but rather represents the (distal) goal at a higher level, i.e. only specifies the (ultimate) goal of the action. For instance, planning to lift a cup does not involve planning the appropriate grasping movement, e.g. moving one's thumb towards their middle finger or lifting one's arm and moving it forward. Rather planning the action in question involves modulation of parts of the subject's sensory and motor systems in order to grasp the cup in front of her.

Each action plan consists of several feature codes, with each code modulating a particular aspect of sensorimotor coordination. It is worth noting here that the same code may also bias other motor systems. For instance, a given code, which controls the 'leftness' of actions by influencing motoric activity in the hand system and drives the hand leftwards, might also influence the eye-system to turn the eyes to the left or the speech system to utter the word 'left' (Hommel et al. 2001, p.

862).[4] Furthermore, a given action has a number of features some of which are more relevant under different circumstances. Which features are relevant on a specific occasion is 'decided'/'selected' in terms of attention, intentions and contextual features.

Even though event codes that ground perception and action planning are fundamentally similar, there are still some differences between the two. For instance, consider a case in which a subject moves her finger upwards and another where the subject's same finger is lifted by another agent. The efferent (or motoric) contribution in each of those two events will grossly differ, while the afferent (or sensory) information will be similar, if not identical. On these grounds, TEC theorists argue that motoric and sensory codes will overlap. Note that even though TEC theorists admit that there are no fundamental differences between the coding of representations that ground perception and action planning, they still allow space for differentiating between stimulus- and response-representations by the role they play in cognitive processes. For instance, representations of the afferent information play different roles in the aforementioned passive and active cases respectively. In the first case (active), they are action or response representations, while in the second case they are stimulus representations, in the sense of representing the passive stance of the subject to an externally controlled stimulus. What is crucial here for TEC theorists is not the coding of the representation but whether the subject is (active condition) or is not (passive condition) in control of the event in question. Thus, the role of an event and the type of its cognitive code is clearly distinguished. It should thus be clear that not only different representations could play the same role but also a given representational code could feature in different functional roles.

27.3.1 Perception Is Not Passive

According to TEC, perception is not a passive process of merely receiving information about the world. Rather perception is a process via which information about the perceiver-environment relationship is processed. This processing includes eye, hand, feet, and body movements, as well as particular allocation of attention, and other cognitive resources. In this sense, we acquire information about the world in an active manner and, in turn, perception is an active process. Crucially, action would ultimately run blind without having perceptual information about its bodily and environmental preconditions, its progress and consequences. In this sense, P&A interact strongly. For instance, the same representation grounds both perceiving a ladder and having the intention to climb up by using the ladder.

[4]Mechanisms that also contribute in solving the binding problem for the representation of perceptual objects coordinate different feature codes. That is, these mechanisms bind together sensory representations of a given object's aspects/parts/features that a subject has selectively attended to and thus represents in a fragmented fashion.

More specifically, the ways in which perception and action planning interact is that both perception and action codes represent the stimulus and the result of a particular sensorimotor coordination. TEC theorists treat anticipation of a perceptual event and planning of an action as aspects of the same cognitive process. In this sense, performing a voluntary action is cognitively identical to anticipating the action's distal effects. Action-generated distal effects are coded and stored together in a common representational code.

A key point in TEC is that the cognitive codes that represent perceptual objects are identical to those representing action plans. The reason for this is that both of these kinds of codes refer to distal events. Nevertheless, the exact sensory code representing a particular spatial distance and the motoric information deployed in order to extend a limb over this distance are not necessarily similar. Instead, what TEC theorists argue is that the representational overlap between sensory and action codes are founded at the more abstract level of the description of the subject's surroundings. Thus, TEC theorists focus on a more abstract distal-coding level. That is, distance, size and location of a given stimulus and response only match in terms of a description of the environmental layout, and crucially not in terms of the specific neural activation patterns that represent it. The sensory code representing a certain distance and the motoric code responsible for driving the hand over the same distance are either matching or mismatching at an abstract distal-coding level. This level is in turn more abstract than the proximal-coding level or coding information about specific arm movements necessary for executing a fetching action, for instance. TEC theorists treat perception and action planning as abstracted from domain- and modality- specific coding. For this reason, TEC theorists focus on event coding.

Perceived and produced events are cognitively represented not as unitary wholes but as bundles of feature codes. Crucially, there is a significant feature overlap between perceived and produced events (i.e. P&A planning). Task relevance and task-related intentions modulate activation and integration of feature codes. Thus, some codes or parts of codes are involved in both perception and action planning. In turn, mutual access to or use of these codes brings about characteristic patterns of facilitation and interference between P&A, and a functional linkage between perception and action planning obtains.

As mentioned in the beginning of the paper, Grounded Cognition theorists argue that sensorimotor representations ground cognition. TEC does provide a plausible story about the nature of sensorimotor representations. However, as explained below (Sect. 27.5.3) this story entails certain compromises for Grounded Cognition theories. In brief, the main options for GCTs are illustrated in Table 27.1 below. Before doing that though, we examine Barsalou's Simulation Theory as a further candidate theory about the ways in which P&A interact and in turn a further view about the nature of sensorimotor representations.

Table 27.1 Levels & implications for GCTs

Level of P&A convergence	Options for GCTs	Implications for GCTs
Higher-cognitive level	(a) Theory of event coding (b) Associationism	Water down "Grounding"
Lower-neuronal level	Associationism	No negative implications

27.4 Barsalou's 'Simulation Theory'

According to Barsalou (1999), thinking is analogous to perceiving. Elaborating on this, Barsalou's starting point is that activation in both sensory and motor parts of the brain grounds perceptual experience. Co-activation of these brain areas yields associations between them. On recalling the object in question, the same representations that were formed during the perceptual experience become reactivated. Thus, thinking is analogous to perceiving to the extent that the same neuronal configurations that were active while *perceiving* a dog, for instance, will also become activated when *thinking* of a dog. The brain *simulates* a given perceptual experience when its cognitive counterpart, i.e. the respective concept (or simulator in ST terminology), becomes activated. Briefly, the claim that concepts are grounded in sensorimotor representations is explicated in terms of concepts becoming activated *in virtue of* activating the appropriate sensorimotor representations. It is worth clarifying that thinking is not identical to perceiving. For unlike perceptual experience, thinking brings about only a quasi-sensory phenomenology.

In illustrating the notion of simulation further, we briefly appeal to Damasio's (1989) 'Convergence Zones' hypothesis. According to Damasio, different neuronal ensembles underlie (or ground, to paraphrase Damasio) perception of different parts/properties of a given object. Further down the line of interneural signaling, the output of the neurons that ground perception of a dog's head, for instance, converge with the output of the neurons that underlie perception of the dog's bark, legs, fur, etc. In this way, these different neuronal ensembles interact in a way that they did not before. The reason why they did not interact before is that they are dedicated to perception of different kinds of stimuli (e.g. auditory instead of visual stimuli, stimuli at different points of the visual field of the subject). Convergence zones register combinations of parts of a given object in terms of coincidence or sequence in space and time (co-occurrence). Reconstruction of the representations of the parts of the perceived object occurs in virtue of 'time-locked retro-activation' of the fragmented records in multiple cortical regions as a result of feedback activity from convergence zones. That is, the groups of neurons that fired in a specific way during the sensory experience with the given object are re-activated simultaneously and in exactly the same way that they were activated during perception of the object in question.

It is worth clarifying that we only have conscious access to the level of a convergence zone and not to the fragmented representations of an object in

geographically spread neuronal ensembles. This is the reason why we perceive objects as wholes and not as collections of parts, features and properties, and why objects are represented as wholes in memory.

27.4.1 Empirical Evidence for Simulation Theory

As already mentioned, both of the candidate theories about the ways that P&A interact are speculative. However, in this section we look into empirical evidence in support of ST in the interest of further motivating our choice to examine ST as well as to flesh out further the nature of sensorimotor representations. In doing so, we look into evidence showing that specific sensory representations get reactivated during imagery. Let us elaborate.

It is widely accepted that quasi-sensory representations are involved in imagery. Furthermore, most would accept that these quasi-representations are formerly formed real sensory representations that become reactivated. The motivation behind this claim stems from the fact that the saccadic movements during perception and during imagery of a given object are too similar to explain it any other way. In this sense, the evidence reviewed below is no *direct* evidence that the same representations formed during perceptual experiences are reactivated during imagery. However, this evidence seems suggestive of the main claim behind ST.

More specifically, Brandt and Stark (1997) show that eye movements recorded while imagining a given object or diagram, are closely related to the ones recorded while viewing the same object or diagram. In fact, the firing patterns of oculomotor cells are associated with sensory cells (since eyes move while perceiving a given object). In order now for the eyes to follow the same or closely matching scanpaths between viewing and imagery it seems plausible to assume that the same oculomotor cells fire. Activation of sensory cells during imagery cause this firing of oculomotor cells. Given the similarities of eye movements during viewing and imagery, it seems that the same sensory cells that underlay perception of the diagram in question were reactivated. Thus, it is in virtue of the associations between sensory and oculomotor cells that the subjects' eyes follow the same or closely matching scanpaths between perception and imagery. These claims also resonate Chao, Haxby and Martin (1999), and Norton and Stark's (1971) 'Scanpath Theory'.

Furthermore, there is evidence that the same neuronal configurations are reactivated also in a bottom-up manner, i.e. on perception of subsequent instances of a given kind, (and not only endogenously, i.e. while thinking of a given object). In turn, this evidence – reviewed below – is seen as suggestive of a convergence between P&A at the lower level.

Demarais and Cohen (1998) focus on whether the nature of a visual imagery (required by a pre-recorded task read to subjects) evokes saccadic eye movements. The obtained evidence shows that eye movements do occur during tasks that evoke spatially extended imagery, and crucially the occurring eye movements reflect the spatial orientation of the image. It is worth noting that subjects activated stored

visual representations of objects featuring in the pre-recorded task, in virtue of an exogenous stimulus reaching the mind through a different modal channel, i.e. auditory vs. visual. Furthermore, the saccadic movements recorded during these experiments suggest that early visual representations and not abstracted conceptual representations are reactivated. In a similar fashion, Spivey and Geng (2001) argue that interpreting a linguistic description of a visual scene requires activation of a spatial mental representation. See also, Farah (1995, 1989), Finke (1989), and Kosslyn et al. (1995), for similar results.

27.4.2 Key Points

One of the main points of ST that is crucial for present purposes is that according to ST, P&A do interact, but one is not constitutively involved in the functioning of the other (in the sense that one could function without the other). P&A are associated in virtue of co-activation of sensory and motor parts of brain, as explained above. Thus, associations between activated sensory and motoric brain regions occur at the lower neuronal level (see below). It is worth clarifying at this point that according to simulation theory, P&A also get 'converged' at the cognitive level in virtue of associations between concepts. This is similar to the TEC claim that P&A 'converge' at the (higher) cognitive level. Despite this similarity though, and unlike TEC, ST does not require a common coding between sensory and motor representations. According to ST, associations between different representational codes (sensory and motoric) suffice to account for the ways in which P&A interact with each other, as well as their relation to cognition. As shown next, this has significant implications for Grounded Cognition projects.

27.5 Theories of Sensorimotor Representations and Grounding

As explained in the previous pages, there is no conclusive evidence for either a constitutive relation (in the sense that you cannot have one without the other) or a functional dichotomy holding between P&A. However, it is widely accepted that there is a strong causal interaction (in the sense that there are contingent causal interactions or influences) between the two. For instance, recall from above Findlay and Gilchrist's (2003) claims about the significance of saccadic eye movements in visual perception. Furthermore, perception of a given object invokes activation of certain motoric representations. For instance, Tucker and Ellis (1998) show that perceiving a cup handle activates the motoric counterparts of a grasping movement. Finally, work on mirror neurons and simulation processes shows that perception is our guide to action (Barsalou 2008).

Given the significant causal relation between P&A, and despite lack of conclusive evidence for a constitutive relation holding between P&A, one could still plausibly talk about *sensorimotor* representations. However, existence of sensorimotor representations does not suffice to secure the Grounding Cognition research programme. For evaluating the implications of the nature of the relation between P&A for Grounded Cognition theories is a complex issue. This complication stems from the fact that there are two points that intersect here:

- The nature of the relation between P&A.
- The point at which P&A intersect. That is, given that P&A do interact, it is only plausible to assume that there is a computational level (either a higher computational level involving concepts or a lower computational level involving only sensory and motoric representations) at which the two converge.

Different answers to the above two issues have different implications for grounding cognition in general and the notion of grounding in particular. Before elaborating on these two issues though, let us briefly digress to make a general remark concerning empiricist theories and our argument.

In this paper, we are focusing on concepts that refer to tangible entities. Thus, our focus is on relatively 'simple' cases for an empiricist program. The reason for focusing on simple cases is that our main concern in this paper is the relation between perception and action, and the implications of this relation GCTs. So, if the problems we analyze in this paper arise for cases of simple concepts, then the same problems can be safely expected to arise in more severe form in cases of complex or lofty concepts. However, we would like to shortly refer to two research strategies that try to account for lofty concepts within an empiricist and GCTs paradigm. First, Wilson-Mendenhall et al. (2013) compared brain activation levels while processing concrete and abstract or lofty concepts and observed great similarities between them. They argue that the meaning of abstract concepts is represented in distributed neural patterns of relevant nonlinguistic semantic content. Reported findings show that the meanings of abstract concepts arise from distributed neural systems that represent concept-specific content.

More specifically, comparison of whole-brain fMRI images obtained while processing abstract concepts like CONVINCE and ARITHMETIC versus concrete concepts like ROLLING and RED, show very few regions emerging. Processing of abstract concepts yielded activations in posterior cingulate, precuneus, right parahippocampal gyrus, and bilateral lingual gyrus. However, processing of concrete concepts yielded a left-lateralized activity profile in middle/inferior frontal gyrus, inferior temporal gyrus, and inferior parietal cortex (p. 930). Interestingly, no activations were observed in the left hemisphere – normally associated with linguistic processing – during processing abstract concepts. This contrasts standard evidence suggesting that abstract concepts are linguistically implemented.

Wilson-Mendenhall et al. report that distributed patterns of activation for the abstract concepts, CONVINCE and ARITHMETIC, occurred in brain areas representing relevant nonlinguistic semantic content. In particular, brain regions implicated in mentalizing and social cognition were active when participants processed

the meaning of CONVINCE. Brain regions associated with numerical cognition were active when participants processed the meaning of ARITHMETIC. These results suggest that abstract concepts are represented by distributed neural patterns that reflect their semantic content, consistent with research on concrete concepts. However, the semantic content unique to different abstract concepts is only revealed when individual concepts are processed deeply in context. Thus, unlike traditional views arguing for linguistic processing of abstract concepts, abstract concepts are treated here as contextually based.

Even though this evidence might not fully vindicate grounding of abstract concepts, the rationale behind interpretations of empirical evidence similar to the above is that by showing that abstract concepts are processed in/by the brain in a manner similar to the one that concrete concepts are processed, GC theorists could resist traditional claims about abstract concepts being processed in virtue of activating amodal symbols. In turn, they can resist the claim that empiricist views cannot account for these kinds of concepts, and thus create some vital space for GCTs.

In a rather more 'traditional' line, Prinz (2002) focuses on the importance of language and suggests that we acquire lofty concepts in virtue of learning the appropriate word. In this way, it is argued that learning abstract concepts is ultimately grounded in perceptual representations (of words). More specifically, Prinz adopts a hybrid semantic account according to which in order for a given concept C to refer to Xs, C has to nomologically covary with Xs, and an X must have been an incipient cause of C. Given that the incipient cause in the case of lofty concepts is an instance of a word, Prinz's account seems to yield the wrong results, i.e. the subject that has acquired the concept democracy, has not acquired the concept of the property but rather the concept of the word. Arguing for Prinz's position lies beyond the scope of this paper, but see Tillas (2010) for a suggestion.

27.5.1 Assuming a Constitutive Relation Between Perception and Action

One of the theoretically possible outcomes of the debate about the nature of the relation between P&A is that a constitution relation holds between the two, in the sense that you cannot have one without the other. We start our analysis by assuming a constitutive relation between P&A. What would it mean for grounding cognition projects, should this assumption be true?

Recall from above that according to grounded cognition theories, sensorimotor representations ground concepts and provide the building blocks for them. In this sense, a given concept is tokened (activated) in virtue of sensory representations being activated. If action is constitutively related to perception, then it is plausible to assume that a set of motor representations will always also become activated alongside the sensory representations, which ground the concept in question. For instance, consider tokening the concept CAT. It necessarily involves activation

of perceptual (visual, auditory, tactile, etc.) representations of experiences with cats. If a constitution relation holds between P&A, then activation of CAT will also involve activation of motoric representations, e.g. representations carrying information about bodily movements involved in stroking a cat sat on your lap, the amount of effort needed for lifting an average sized cat, etc.

In order for Grounded Cognition theories to be successful, different sensory and motoric activation patterns have to ground the meaning of each concept; this is clearly problematic. For there seems to be a huge difference between representations at the conceptual and the motoric level, regarding their fine-grainedness. On the one hand, motoric representations seem to be very specific, for instance representations used in making a very delicate movement. On the other hand though, the same representations could be used in making various movements captured by different action concepts (or verbs). Consider for instance the concept GRASP (or the verb 'to grasp'), which refers to grasping movements in general. Grasping movements can be very different in terms of their motoric activation patterns. At first sight there seems to be a mismatch at the level of fine-grainedness between representations of specific grasping movements and representations that would refer to all possible grasping movements, e.g. those representations comprising the concept GRASP. In order to settle this issue, one has to decide about the most appropriate level of abstractness of motoric representations, and whether abstracted representations could legitimately qualify as motoric. A possible way to deal with this discrepancy between representations at the conceptual and motoric level is to assume that there is a certain level of abstractness already at the motoric level. In our reading, this is what TEC theorists seem to imply while elaborating on the processing of action-planning-related representations. Whether this is a plausible claim to make and the extent to which it enjoys empirical support remains debatable. It is also worth clarifying that there is a tension between the claim that tokening of concepts occurs in virtue of *abstracted* motor representations and evidence showing that representations of *particular* perceptual experiences become reactivated during imagery (see Sect. 27.4 above).

27.5.2 Where Do P&A Meet?

As explained above, P&A converge or meet in two ways. First, it is clear that there are neuronal ensembles dedicated to sensory perception and neuronal ensembles dedicated to processing of motoric activations. The point where P&A meet is the point of interneural signaling where the neuronal ensembles converge. Second, P&A could meet in terms of associations between sensory and motoric representations. These two possible convergence points between P&A imply that there are two possible levels at which this converge could occur:

(a) A lower level of pre-conceptual processing.
(b) A higher level of conceptual (cognitive) processing.

Regardless of the nature of the relation between P&A there will always be (a) brain region(s) that realise this convergence. Nevertheless, what is crucial is that different views about the relation between P&A and their level of convergence have different implications for the notion of grounding. Let us run the two points together, and elaborate further on the implications for GCTs.

Assuming a constitution relation holding between P&A (in the sense that you cannot have one without the other), there are two options to explain the level at which the two converge, as well as the way in which this is done.

27.5.3 Convergence Under Constitution: Meeting at a Higher Level

One possibility is that P&A converge at a higher cognitive computational level. In the light of the candidate theories we examined here, this can occur in two ways:

(i) Theory of Event Coding: On perceiving a given object, the late output of the perceptual experience in question will be *necessarily* (since a constitution relation is assumed between P&A) fed into the process where the early antecedents of related action planning occur. (As already mentioned, according to TEC theorists, the output of the first stage and the input of the second stage are commonly coded).

(ii) Simulation Theory (& Associationism): Assuming a constitution relation between P&A, on thinking of a given object the brain simulates both the perceptual and motoric states in which a given subject was during perception of the object in question.

In our reading, TEC suggests a plausible way in which sensory and motor representations are coded. However, TEC and GCTs cannot happily coexist, since TEC entail a series of negative implications for GCTs. In particular, the problem for GCTs is that TEC's starting point is already at a cognitive level (cf. Sect. 27.3), such that GC theorists have to choose between the following two options:

1. Weaken the notion of grounding: According to the working hypothesis of this paper, grounding means 'analyzing' something in terms of something else, which resides at a slightly lower level. If GC theorists couch the notion of grounding in terms of TEC, and given that both concepts and event codes are at the same (cognitive) level, the notion of grounding seems compromised. For the target of GCTs is to ground cognition in the lower level where sensorimotor representations reside, and TEC concerns representations at a higher level.
2. TEC as the first step of grounding: Another option for GCTs is to argue that TEC only specifies an intermediate level (early cognitive level) where sensory and motor representations are coded together. In turn, sensory and motoric representations ground representations at this intermediate level. In this way, GC theorists can avoid the aforementioned compromises. However, the gains from this move are far from clear. For proponents of GCTs will still have to

explain the relation between sensory and motor representations prior to their commonly coding. Ignoring potential further difficulties for now, there are again two options for grounded cognition theorists. First, GCTs could still argue that tokening of a concept occurs in virtue of activations of sensory and motoric representations, and that at some point of interneural processing, those two kinds of representations are commonly coded. In this case, TEC is of a little help for GCTs. An alternative is to argue that once sensory and motor representations become commonly coded, the antecedents of this commonly coded composite sensorimotor representation becomes idle and are not involved in cognitive processing. Thus, thinking occurs in virtue of reactivating the 'composite' or commonly coded sensorimotor representation – to put it in TEC terms. It is worth stressing that in the latter case, GCTs avoid the aforementioned compromise. For it might well be the case that sensory and motoric representations at the lower level ground concepts. It is just that tokening of concepts occurs in virtue of merely activating the commonly coded composite representation, and not its antecedent sensory and motor representations. However, evidence for scanpath theory (see above) seems to potentially undermine the idle-role-of-low-level-representations claim. For this evidence suggests that low-level sensory (quasi-)representations are reactivated during imagery. If proponents of GCTs explain tokening of concepts in terms of activation of the TEC-like representation, then they should find a way to accommodate the above evidence. Given that there is not a perfect match between scanpaths during perception and imagery, GC theorists could argue that this is due to activation of abstract commonly coded representations, while sensorimotor representations of particular experiences remain idle. A further alternative for GCTs is to argue that representations at the TEC level do carry enough information initially captured in terms of sensory and motoric representations. In this way, GCTs could explain the similarities between scanpaths recorded during perception and imagery of the same object/event.

It is worth clarifying that the above analysis of TEC might be in tension with GCTs. First, it allows for an understanding of the commonly coded representations as amodal, which is in direct contrast to GCTs. Second, on the above construal TEC seems in line with 'Classical Sandwich' views, since P&A meet at the cognitive level. Once again, this is in contrast to the main claims of GCTs about *direct* interaction between P&A.

In reply to those potential criticisms, TEC actually proposes a common coding between *late* cognitive perceptual products and *early* cognitive action *antecedents*. To this extent, the commonly coded representations are clearly of sensorimotor nature, and P&A do interact directly even if this 'convergence' at the *early* cognitive level. To highlight the contrast between TEC and Classical Sandwich views further, consider that according to the latter, the output of perceptual processes is transduced into an amodal code, which is then fed into action processes. Therefore, TEC seems immune to the above criticisms. Nevertheless, as already shown, TEC is still probably not the best option for GCTs given the aforementioned compromises.

The only way that TEC could be useful for GCTs is if GC theorists argue that the common coding between sensory and motoric representations is the first step in the grounding process (starting from the higher level). GCTs could then appeal to Associationism (see below) to account for the relation between P&A at the lower level. In this case though, the common coding first step is somewhat redundant – at least for purposes of grounding – and it needlessly complicates things. That said, it does no further harm, and it might be a good option for GCTs to adopt this 'first-step-strategy' especially given that TEC enjoys strong support from independent empirical evidence.

In brief, TEC seems to suggest a plausible way in which P&A are related. However, there are certain 'negative' implications for grounded cognition views. Next, we turn to examine whether the relation between P&A could be accounted for at a lower level, to the one TEC suggests, and whether the aforementioned concerns could be avoided.

27.5.4 Convergence Under Constitution: Meeting at the Lower Level

In this subsection, we examine the implications that a constitutive relation between P&A would have for GCTs. Recall from above that a constitutive relation between P&A is most often taken to mean that you cannot have one without the other and vice versa.[5] However, even if there is a system that can solely perceive but not execute (actions), and vice versa, one can still plausibly talk about sensorimotor representations. For sensory and motor representations could be associatively (and not constitutively) related.

Furthermore, even if P&A are not constitutively related, i.e. they are not constitutive for each other, they can still be constitutive for something else, namely sensorimotor representations. That is, it is plausible to assume that there are sensory and motoric representations that constitute sensorimotor representations, without implying that P&A are themselves constitutively related. In this sense, the agenda of Grounded Cognition theorists is not dependent on a constitutive relation between P&A.

Once again, we start by assuming a constitutive relation between P&A. Given that in this section we examine the possibility of a convergence between P&A at the lower level, we consider an additional factor. Namely, we consider the intuitive expectation that there is an interface between P&A. This interface is/(are) (a) brain region(s) that realizes the sensorimotor representation in question. Finally, we take into account the – clearly empirical – hypothesis behind GCTs according to which activation of concepts occurs in virtue of activation of some brain regions. In the

[5]However, recall that there are other ways to understand the relation between P&A; for instance that the two are functionally distinct and that one could operate without the other, as shown below.

interest of illustrating this point further, consider a specific example of a P&A-interface such as the 'visual-and-motor neurons' found in the monkey's parietal cortex, and the 'mirror neurons' located in the premotor cortex areas that are commonly associated with action planning. Most consider the mirror neuron system (MNS) to be a valid candidate for an interface between P&A. *If* the MNS was the *only* interface between P&A – which is clearly not the case –, and if GCTs are correct, then this would imply that MNS are involved in processing of all concepts. This would in turn be clearly problematic. For even though it is plausible to assume that the MNS contributes to cognitive functioning – for instance mimicry is crucial for social cognition and behavior – it is absurd to expect activation in the MNS during all cognitive processes. Therefore, the MNS cannot play this role. As an offshoot, *if* the MNS was the *only* interface between perception and action, then GCTs and proponents of the claim that a constitutive relation holds between P&A would seem mutually exclusive.

A further example of brain regions that could serve as the interface between P&A comes from Gallese (2007). In particular, Gallese argues that the ventro-dorsal stream, which involves projections from the inferior parietal lobe to the pre-frontal and pre-motor areas, serves as the main interaction zone for the ventral and dorsal streams, which are in turn seen as vision for perception and vision for action respectively.

If a constitutive relation holds between P&A, and at the same time sensorimotor representations ground cognition, then one would expect activation of parts of the ventro-dorsal system when tokening concepts. In fact, the more successful grounded cognition theories are, the more plausible it would seem to expect activation in those brain regions when tokening concepts. Once again, and *if* the ventro-dorsal system is seen as the *only* interface between P&A, there seems to be a tension between a constitutive relation between P&A and GCTs.

Assuming a single interface between P&A as well as a constitutive relation between the two, what are the implications for GCTs? Given the aforementioned tension, and *if* the P&A interface is a single part of the brain, then it seems preferable to drop the claim about a constitution relation holding between P&A rather than the Grounded Cognition hypothesis. For the former potentially entails evolutionary concerns such as the aforementioned tokening of the MNS during all cognitive tasks. Crucially, and as shown already, a constitution relation is not in principle incompatible with associationism and thus does not imply a 'single' interface between P&A (hence the above conditional). Having a collection of geographically spread associated brain regions as an interface between P&A is a further option for GCTs – one we examine next.

27.6 What If P&A Are Functionally Distinct?

In contrast to the working hypothesis of the previous section, we now consider the possibility that P&A are functionally distinct. As mentioned already, the agenda of GCTs does not depend on a constitutive relation between (P&A) – in the sense that

Table 27.2 Levels and implications for GCTs

Level of P&A convergence	Option for GCTs	Implications for GCTs	P&A relation		P&A interface	
			Constitution	*f*-Dichotomy	Single	Spread-out
Higher-cognitive level	(1) TEC (2) Associationism	Water down "Grounding"	*Necessary* common coding OR formation of associations between S&M reps.	Common coding OR formation of associations between S&M reps.	Evolutionary concerns	(a) Intuitive (b) Empirically vindicated (c) No evolutionally concerns
Lower-neuronal level	Associationism	No negative implications	*Necessary* formation of associations	Formation of associations between S&M reps		

you cannot have the one without the other. In turn, a constitutive relation between P&A is not a (necessary) precondition for sensorimotor representations to qualify as legitimate.

If P&A are functionally distinct, what are the implications for GCTs? More specifically, what are the implications of a functional dichotomy between P&A for *sensorimotor* representations? Once again, there are two possible levels at which sensory and motor representations could 'converge'.

(a) Convergence at a lower neuronal level: Representations that are sensory and motoric in nature get associated in virtue of co-occurrence during perceptual experiences. By appealing to associationism, it is plausible to assume that sensorimotor representations ground concepts, even if P&A are functionally distinct.
(b) Convergence at the cognitive level: Concepts could be grounded in a high level, similar to the one suggested by TEC theorists (see also Sect. 27.5.3). Associations between sensory and motoric representations could also occur at this higher level.

In brief, the above issues are illustrated in Table 27.2 above.

27.7 Take Home Message

From the above analysis, it should be clear that there is currently no conclusive evidence either for a constitutive relation holding between P&A (in the sense that you cannot have one without the other) or for a functional independence between the two. At the same time, there is no conclusive empirical evidence for grounding cognition either. Thus, at present one could only put forth some speculative hypotheses about the implications of the outcome of the debate about the nature of the relation between P&A and the issue of grounding concepts in sensorimotor

representations. In an attempt to put forth the most plausible hypothesis with the least negative theoretical implications for GCTs, we considered the following options:

(a) TEC provides good reasons to assume that P&A converge at a higher level. However, and unlike associationism, this entails a compromise for grounded cognition theories.
(b) Regardless of the outcome of the debate about the nature of the relation between P&A, associated sensory and motor representations could ground concepts.
(c) Associationism is a minimum requirement thesis to the extent that:

- It does not imply the need for a common coding system.
- It does not yield evolutionary concerns (e.g. the mirror neuron system being involved in cognition).[6]
- Associations could occur at both a lower and a higher level.

From the above, we argue that talking about '*sensorimotor*' representations is indeed legitimate. In turn, given that one of the central claims behind GCTs is that sensorimotor representations ground concepts, this paper creates some vital space for GCTs. Crucially, this vital space does not depend on the outcome of the debate about the nature of the relation holding between P&A.

Acknowledgements We would like to thank Patrice Soom, James Trafford, and Uwe Peters for their help and comments on earlier drafts.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–609.
Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.
Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience, 9*, 27–38.
Catmur, C., Walsh, V., & Heyes, C. (2007). Sensorimotor learning configures the human mirror system. *Current Biology, 17*, 1527–1531.
Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2*, 913–919.
Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*, 25–62.
Demarais, A. M., & Cohen, B. H. (1998). Evidence for image-scanning eye movements during transitive inference. *Biological Psychology, 49*(3), 229–247.
Farah, M. J. (1989). The neuropsychology of mental imagery. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 2, pp. 239–248). Amsterdam: Elsevier.
Farah, M. J. (1995). Current issues in the neuropsychology of image generation. *Neuropsychologia, 33*, 1455–1471.
Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.
Finke, R. A. (1989). *Principles of mental imagery*. Cambridge, MA: MIT Press.

[6]Note here that there is evidence (Catmur et al. 2007) that mirror neurons could be associationistically trained.

Gallese, V. (2007). The 'conscious' dorsal stream: Embodied simulation and its role in space and action conscious awareness. *Psyche*, *13*(1) (archived electronic journal: http://psyche.cs.monash.edu.au/)

Gangopadhyay, N., Madary, M., & Spicer, F. (2010). *Perception, action and consciousness*. New York: Oxford University Press.

Goodale, M. A., & Milner, A. D. (2004). *Sight unseen: An exploration of conscious and unconscious vision*. Oxford: Oxford University Press.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*(5), 849–878.

Horst, S. (2009). The computational theory of mind. In *Stanford encyclopedia of philosophy entry*.http://plato.stanford.edu/entries/computational-mind/

Hurley, S. (2001). Perception and action: Alternative views. *Synthese, 129*, 3–40.

Jacob, P., & de Vignemont, F. (2010). Spatial coordinates and phenomenology in the two-visual systems model. In N. Gangopadhyay, M. Madary, & F. Spicer (Eds.), *Perception, action and consciousness* (pp. 125–144). Oxford: Oxford University Press.

Jacob, P., & Jeannerod, M. (2003). *Ways of seeing, the scope and limits of visual cognition*. Oxford: Oxford University Press.

Kiverstein, J. (2010). Sensorimotor knowledge and the contents of experience. In N. Gangopdhay, M. Madary, & F. Spicer (Eds.), *Perception, action and consciousness: Sensorimotor dynamics and dual vision* (pp. 257–275). New York: Oxford University Press.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature, 378*, 496–498.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. Oxford/New York: Oxford University Press.

Milner, A. D., & Goodale, M. A. (2010). Cortical visual systems for perception and action. In N. Gangopadhyay, M. Madary, & F. Spicer (Eds.), *Perception, action and consciousness*. New York: Oxford University Press.

Noë, A. (2005). *Action in perception* (pp. 71–94). Cambridge, MA: MIT Press.

Norton, D., & Stark, L. W. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research, 11*, 929–942.

O'Regan, K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 883–917.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129–154.

Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT Press.

Shapiro, L. (2007). The embodied cognition research programme. *Philosophy Compass*. Article first published online 2 feb 2007. doi:10.1111/j.1747-9991.2007.00064.x.

Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research/Psychologische Forschung, 65*(4), 235–241.

Tillas, A. (2010). *Back to out senses: An empiricist on concept acquisition*. Ph.D. thesis. University of Bristol.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology, 24*(3), 830–846.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. W. A. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

Weber, A., & Vosgerau, G. (2012). Grounding action representations. *Review of Philosophy and Psychology, 3*, 53–69.

Wilson, M. (2002). Six views of embodied cognition. *Psychological Bulletin and Review, 9*, 625–636.

Wilson-Mendenhall, C. D., Simmons, W. K., Martin, A., & Barsalou, L. W. (2013). Contextual processing of abstract concepts reveals neural representations of nonlinguistic semantic content. *Journal of Cognitive Neuroscience, 25*(6), 920–935. doi:10.1162/jocn_a_00361.

Chapter 28
The Seminal Speculation of a Precursor: Elements of Embodied Cognition and Situated AI in Alan Turing

Massimiliano L. Cappuccio

Abstract Turing's visionary contribution to cognitive science is not limited to the foundation of the symbolist approach to cognition and to the exploration of the connectionist approach: it additionally anticipated the germinal disclosure of the embodied approach. Even if Turing never directly dealt with the foundational speculation on the conceptual premises of embodiment, in his theoretical papers we find traces of the idea that a cognitive agent must develop a history of coupling with its natural and social environment, and that primitive bodily stimuli like pain and pleasure drive this coupling and elevate it to real learning by setting its normative preconditions. Turing did not consistently defend the centrality of embodiment, and ended up confounding or deemphasizing in various occasions the critical importance that he had himself implicitly recognized to the body. In line with the anti-representationist, radically enactive approaches to basic cognition, I believe that if Turing eventually failed to fully value the cognitive-developmental role played by the body, this was not because he proposed a computational and functionalist model of the mind, but because he tacitly assumed the content/vehicle dichotomy as a primitive of that model: in fact, he still believed that intelligence is a realized by decontextualized contents that can be detached and transmitted regardless of their mode of physical implementation.

Keywords AI • Alan turing • Developmental robotics • Embodied cognition • Connectionism • Cognitivism

28.1 Turing's Legacy and Cog Sci

Alan Mathison Turing is regarded as one of the fathers of computer science, and since computers represented for decades the main paradigm for the sciences of the mind (Pinker 2002; Boden 2006), his name is inextricably tied to the very

M.L. Cappuccio (✉)
United Arab Emirates University, Al Ain, UAE
e-mail: massimilianocappuccio@hotmail.com

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_28

birth of cognitive science as a research program. To evaluate the importance of this legacy, one must situate it in a historical perspective, considering the three prototypical approaches that (according to Varela et al. 1991) inspired cognitive science through its evolution. First, the cognitivist – or classical – approach, which is logico-symbolic in nature and based on the idea that a cognitive system is centrally realized by the sequential manipulation of discrete units of information in accord with formal rules of combination and substitution (Miller 1956; Neisser 1967; Minsky 1974; Haugeland 1978). Second, the connectionist approach, tied to the idea that information, realized by super positional and distributed representations, is stored and manipulated through parallel systems implemented by neural networks (McCulloch and Pitts 1943; Rumelhart et al. 1986; van Gelder 1991). Finally, the embodied approach (Varela et al. 1991; Clark 1997; Gallagher 2005), stressing that the details of the material, organismic, and experiential implementation of bodily skills and behaviors play an active causal role in structuring cognitive functions (for example, fine-tuning the sensorimotor capabilities of a system, cfr. Noe and O'Regan 2002).

The first approach defines intelligent decisions as controlled by stored sets of 'rules of thumb' that reflect an internally localized representation of the behavior to be delivered by the system in specific circumstances. In opposition to this, the second approach says that the contents of these representations are memorized through stochastic reinforcement mechanisms when the weights of the network nodes are modified by environmental feedbacks. The third approach maintains, contrary to the first and the second, that the coupling between the system and the environment is not stored as a contentful representation of states of affairs, but a dynamical (Dreyfus 2002; Freeman 2000) and emergent (Petitot et al. 1999) know-how that both modifies and is modified by the system's perceptual and motoric dispositions while adapting to contextual circumstances. Some versions of the embodied approach stress that the cognitive processing responsible for this adaptive capability can be partly outsourced to bodily (non-neural), or even extra-bodily, elements (Clark and Chalmers 1998); others underline the necessary contribution of situated interaction and mutual unprincipled engagement in development and evolution (De Jaegher et al. 2010; Hobson 2002; Reddy 2008).

The history of computer science saw in Turing one of the heroes of the first approach. Turing's computational machine (also called "discrete-state machine", Turing 1948) provides the prototypical model to envision how the algorithms underlying certain intelligent functions can be centrally controlled and implemented by logico-symbolic operations realized through mechanical means. However, while most of his colleagues were busy trying to replicate human intelligent functions through the logico-symbolic machines that he had theorized, Turing was designing connectionist intelligent systems, as he had already started working on some of the earliest models of parallel distributed processing through neural networks (which he called "unorganized machines" in 1948). In this paper, I would like to point out that Turing's visionary contribution to cognitive science is not limited to the foundation of the first and to the exploration of the second approach to cognition, as it additionally anticipated the germinal disclosure of the third approach.

In some of his most speculative works (Turing 1948, 1950), Turing began pondering upon at least some of the key insights that a few decades later inspired the embodied approach: namely, we find traces of the idea that a cognitive agent must develop a history of coupling with its natural and social environment, and that primitive bodily stimuli like pain and pleasure drive this coupling and elevate it to real learning by setting its normative preconditions. Additionally, even if this chapter will not deal with Turing's contribution to theoretical biology, it is important to mention that Turing's views on the development of cognition go on par with his ground-breaking work on morphogenesis (1952), where he characterized living functions as developing from bio-chemical patterns that are not stored or represented as sets of rules, but generated by the stochastic equilibrium of reaction-diffusion processes determined by the fine-grained interplay of organismic predispositions and aleatory environmental circumstances. This is particularly important for Turing's treatment of self-organizing learning machines: as pointed out by Lassègue (1998, p. 113), their "self-organization progressively constitute for him 'the simplest model of a nervous system having a random assemblage of neurons'. There is the outline of what later will become the domain of the 'neural networks' and of their adaptive properties."

Attributing to Turing the paternity of the embodied-embedded approach to cognition would be excessive and misleading, as he never directly dealt with the foundational speculation required by the theoretical premises of embodiment, and he certainly didn't have sufficient time and technological means to explore it empirically. Nonetheless, re-reading his works 100 years after his birth, one realizes that some of the key conceptual points seminally addressed in his papers are the same that, much later, motivated the analyses on the embodied nature of cognition. I would like to try to show that, if Turing was certainly not the father of embodied cognition, he was not a frontrunner of cognitivism either, even if it is true that his work turned out to be instrumental to support the rise of cognitivism.

28.2 Cog Sci and Turing Machine

In 1936, Turing proposed the first speculative model of a "logical computing machine" (later simply referred by others to as "Turing machine") to demonstrate what later would be called "Turing thesis" (or "Church-Turing thesis", Kleene 1952): any computable number or sequence is computable by such a machine. This means that any algorithmic function can be realized through a finite number of operations conducted on discrete symbolic elements and decided by a pre-given set of rules of manipulation, or "tables of behavior", which are both logically (formally) necessary and physically (mechanically) deterministic. But this thesis does not imply that the functioning of the human mind is computational in nature or realized by algorithms. We cannot find traces of this idea in the 1936s paper, which aimed at demonstrating the inverse assertion: each algorithmic process of thinking that can be carried out by a human being through formal procedures can be exactly reproduced by computational and mechanical means. These two assertions are compatible, yet their difference is substantial: the latter describes the powers and the constraints

of the recursive operations that we can effectively use to formalize mathematical processes, and does not imply the former, which is ontologically stronger, and relates to the deep constitution of psychological processes. The latter states that a certain practice (calculus) shapes a certain form of thinking (calculative), and that a key feature of that practice can be mechanized; the former asserts that if that practice is mechanizable it is because the underlying form of thinking is inherently mechanical.

This difference was destined to be overlooked, after Turing, especially because computer science started accumulating striking successes in replicating human formal procedures by means of sophisticated electronic implementations of the Turing machine. Since then, the universal Turing machine has been seen more and more an irresistible metaphor to describe in mechanical terms the fundamental processes that govern the functioning of the human mind, reviving the long-standing rationalist belief that thought is, after all, nothing else than calculus on mental contents (Longo 1999).

Skepticism towards this expectation has been growing especially in the last decades, and today we are better inclined to appreciate the primary and truest sense of this metaphor. Lassègue (1998, p. 202) observed that the great epistemological novelty brought by Turing's machines is "the biological aspect of this characterization of the mechanism: therefore, it is more the machine that resembles the organism than the organism resembling the machine". Ludwig Wittgenstein (1980, § 1096) had already warned us, with his dry comment, long ago: "Turing machine. These are humans who calculate". These comments, stressing the biologic and anthropomorphic background of Turing's idea, invite us to reject the cognitivist interpretation of Turing thesis, i.e. the claims that – after all – human minds are nothing other than calculating machines, and replace it with the inverse, constructivist interpretation: calculating machines – after all – are nothing other than devices built to surrogate a human practice.

The latter interpretation turns out to be not only plausible, but compelling, if we take seriously the fact that Turing thesis, exactly like its equivalent formulations (e.g., lambda-calculus, cfr. Church 1936; recursion theory, Kleene 1936), is not a theorem demonstrable by the power of mechanical calculus, but an intuitive evidence ("a direct appeal to intuition", Turing 1936, p. 249) revealed by the phenomenology of the human practice of calculus. The importance of this phenomenology appears clearly in Sect. 9 of *On computable numbers*, where Turing reveals what his machine was originally meant to be: a stylized illustration of the embodied writing/reading operations that have to be conducted by a "computer" in flesh and bone, when (s)he is busy at a desk to solve some mathematical problem with pencil and paper.

28.3 Turing Machines and Human Computers

Looking closer at how Turing describes the structure of his machine in Sect. 9 of his 1936 paper, we see how he transfigures the bodily elements of the human computational practices, replacing them with their mechanical counterparts: the

bi-dimensional squared paper sheets used by humans as physical supports of computation are substituted by the mono-dimensional tape of the machine, whose only functionally relevant characteristics are its unconstrained linear extension, its double direction of movement, and the fact that it is segmented in a series of equal "sections"; the body of the human computer, in turn, with its organs and appendices dedicated to manipulation and vision, is substituted by the moving head of the machine, whose only functions are sliding on the tape and scanning, typing, or erasing a symbol at once; finally, the attentional, dispositional, and ideational capabilities of the human computer are reduced to the machine's moving head's power to be "directly aware" of a single symbol at once and "keep in mind" the functional "configuration" that determines its following action (p. 231). The explanatory power of this illustration, which offers an abstractly universal, yet fully operative account of the mechanical nature of algorithmic practices, derives precisely from being both extremely concrete and, at the same time, effective in obliterating the original source of such concreteness. In fact, while it clearly shows that the human computer's body is the necessary and constitutive precondition for any computational operation, it concurrently reduces the role of the body to a redundant (functionally irrelevant) material accident of that very operation.

This ambivalence, which preludes to Turing's incomplete investigations of the embodiment of intelligence, is hinted at by the grammatological norms that define the machine's behavior. On the one hand, the operations illustrated by the Turing machine are instantiations of a writing practice, i.e. the simplest embodied acts that a human computer must carry out to calculate, "so elementary that it is not easy to imagine them further divided" (p. 250). As constituents of a graphic procedure, they can be sharply distinguished as either sensory (scanning a series of symbols on a sheet, one at a time) or motoric (erasing the currently scanned symbol or printing a new one, and moving to scan another symbol), their goal being, respectively, to recognize and produce a series of written graphic elements that belong to a well-defined alphabet.

On the other hand, just because the practice of mathematical calculus is analyzed into allegedly atomic graphic manipulations reducible to logical operations in a dimension of virtual necessity, Turing's illustration is stylized and decontextualized. What it captures of the living practice of writing is only its formal features, i.e., the discrete transitions that generally look (to a mathematician) strictly indispensable to account for the logical role played by writing in the creative task of solving a mathematical problem. Anti-formalist mathematicians (Longo 2002, building on Weyl 1985) and psychologists of rational processes (Lakoff and Núñez 2000) stress that there is much more in mathematical writings than blind and empty rule-based mechanical procedures; at the same time, they should recognize that, if such formalistic reduction is virtually possible, it is primarily because it is allowed by certain writing practices whose norms are discrete and mechanistic in nature.

Note: Turing machine is not just a bodiless replication of the body of a living mathematician; it is also an idealized explication of the norms that guide her intelligent skills, transposed into an abstract space of typographic exactness (Cappuccio 2005, 2006; Herrenschmidt 2007; Lassègue and Longo 2012). Turing's

insight into the structural preconditions of the mechanization of computation matches very well with Turing's late reflection, where – among other things – the British polymath investigated the theoretical and practical possibility to reproduce the entirety of human thought by mechanical means. We should, therefore, refer to *Intelligent machinery* (1948) and to *Computing machinery and intelligence* (1950), two influential texts in which Turing's speculative enquiry into the possibility to build machines capable of human-like intelligence is famously addressed. "Can machine's think?", the question that specifically stimulates his 1950 paper confirms that the ambitions of his investigation were still in line with his earlier work on the nature of computation: once again, the question is whether (and how) machine intelligence could one day become so complex to rival human thinking, not whether the processes of human thought are inherently mechanical.

Turing famously refuses to address the question in this ontologically demanding form, which – according to him – typically relies on very speculative and too abstract notions of "thought". Rather, the proposed criterion to evaluate whether intelligence has been achieved or not by machines is entrusted to an "imitation game", a procedure to empirically determine whether a human player's interaction with a machine appears to him indistinguishable from the usual interaction with other humans. Discussing Turing controversial choice on this regard, the commentators often stressed that this method leads to collapse the distinction between truly intelligent decisions and stereotyped routines that mimic their exterior appearances. This reflection also raised well-motivated concerns about some of the strongest explanatory expectations on AI (Searle 1980).

But there is another issue that deserves to be deepened: why do we need a game based on the actual human-machine interaction (whether competitive or cooperative in nature) to test the successful implementation of intelligence by artificial means? Interestingly, what Turing proves in this famous paper is not that in order to reproduce intelligence a machine needs a certain logico-symbolic computational architecture, but that it must be capable of adapting to the phenomenology of a social encounter with the human (Gallagher 2009, 2012). That is why one can't find any ultimate commitment to a logico-symbolical account of human intelligence in Turing's work (Longo 2008). On the contrary, the imitation game shows that a certain behavior looks credibly intelligent to a human observer only if she recognizes it as familiar; and that this familiarity is measured by the degree of interactive attunement that the observer spontaneously tends to establish with the AI through conversation.

But we know that such attunement largely depends on experiencing a corporeal relation of reciprocity (De Jaegher et al. 2010) – a bond that might be mediated by the capability to empathize through one's own body with the body of another (Gallese 2009). At a first glance, in Turing's description of the imitation game the role of the body of the participants (whether human or robotic) seems systematically removed: by design, their physical identity is hidden, and their communication is restricted to perceptually standardized and emotionally neutral exchanges of strings of symbols. The player can only infer the interlocutor's intentions, which can't be directly manifested through bodily expression, unlike in the typical human forms of social engagement. According to Turing, the imitation game is reliable precisely

because it measures the intelligence of the machine by drawing "a fairly sharp line between the physical and the intellectual capacities of a man." He seems convinced that there would be little point in trying to make a "thinking machine" more human by dressing it up in [. . .] artificial flesh", even if "engineers and chemists" were one day able to provide a machine with such bodily features.

This seems irremediably at odds with any interactionist and embodied approach to cognition and to sociality. Nevertheless, I would like to stress that, if Turing pointed to a social interactive context as the most reliable scenario to test the successful development of intelligence, this is because he had clear in his mind that direct, unprincipled embodied forms of intersubjective engagement are constitutive *at least* of the earliest development of their intelligence, and that the early bodily forms of interactive engagement with others are a precondition for the most developed forms of intelligence. Indeed, Turing takes seriously the psychological link between embodied sociality and cognitive performance, though he doesn't thoroughly discuss its implication: he is always tempted to replace the defining role played by the body in its contingent modes of implementation with context-independent heuristics, formal rules, and mechanical procedures; nonetheless, rather than stating that the human thought is entirely reducible to logico-symbolic processes, he carefully describes how, in order to achieve intelligence, even a machine needs a body.

Two facts suggest that Turing was struggling to move beyond the cognitivist framework. In the first place, he never claims that, in order to think, a cognizer necessarily needs to implement a cognitivist cognitive architecture (logico-symbolic, sequential, discrete, centrally controlled, syntactically organized, fully-representational, internally localized); on the contrary, as I mentioned before, in 1948 he had carefully considered the functionality of parallel-distributed connectionist models, and was well aware that brains are "unorganized machines", highly plastic systems sensitive to limit conditions (Longo 2002, p. 3). In the second place, in 1950, Turing claims that, even if human cognitive functions can in principle be *imitated* by logico-symbolic processes (and this is why he keeps considering discrete-state machines as the standard for his game), the *emergence* of these functions (i.e., the developmentally necessary preconditions of their genesis and actual implementability) must be rooted into an embodied process of learning based on situated interactions and unprincipled exploration. Indeed, he admittedly does not have any positive arguments to convince his readers that digital computers can fully imitate human intelligence (p. 443) and, in a constructivist fashion, he can only rely on the persuasiveness of the description of a child-machine as a genetic model of the educational and experiential conditions that could lead a computer to become intelligent. Significantly, in his account, the question into the possibility to recreate human intelligence by mechanical means becomes a question into the possibility to recreate human development by artificial means.

Now, Turing did not consistently defend the centrality of embodiment, and ended up confounding or deemphasizing in various occasions the critical importance that he had himself implicitly recognized to the body. But, this unfinished speculation can't cancel the fact that the 1950 paper is not meant to support a mechanist reduction of intelligence to computation. This point is ignored by his commentators

more often than it should: Turing knew that an intelligent system requires a fine-grained calibration to contextual circumstances, and that only a living organic body can achieve it through a history of complex adaptive interactions with the world and with other agents.

28.4 Human Computers and Child-Machines

Today, contextual circumstances of the embodied/embedded kind are considered key by the most promising approaches to A.I., e.g. situated (Hendriks-Jansen 1996), evolutionary (Nolfi Florean 2000) and developmental robotics (Cangelosi and Schlesinger 2015) which often go hand in hand with the new anti-cognitivist and post-connectionist wave in cog sci: according to these approaches, it is the body's material predisposition to balance the unpredictable fuzziness and massive complexity of the sensorimotor circumstances that establishes implicit and pre-categorical norms for the stability, effectiveness, and flexibility of intelligence; in turn, these norms can inform an environmentally entrenched intelligence by moulding the agent's bodily dispositions to action and perception in the ways that best allow the system to fluidly respond to concrete situations so to reduce its distance from a foreseen optimal balance with the contingencies (Freeman 2000; Dreyfus 2002). An embodied system learns by enacting through its actions complex feedback loops that continuously inform the system of its own sensorial, proprioceptive, and motoric possibilities and constraints (Brooks 1991; Clark 1997). Such body – with its intrinsic adaptive and predictive functions – is what specifically determines the form and the direction of the machine's cognitive processes, as it acts as a distributor, a constraint, and a regulator of the underlying mechanisms of decision and control (Shapiro 2011). But, most importantly, what enables and coordinates these three functions is the fact that the body constitutes the most fundamental background of the sense-making activity of a cognitive system: flexibly adaptive dispositions truly able to realize intelligence can only emerge if they are materially embedded as a tacit (irrepresentable) know-how that is directly responsive to concrete contextual contingencies (Dreyfus 2008). Bodily sensitivity to real contexts is what enables the immediate intelligibility of relevance and the prioritization of the teleological dispositions of the living system.

Turing certainly contemplated the possibility that embodiment could be a tacit precondition of intelligence, but suspected that the success of an embodied AI would be contingent upon substantial future advancements in bio-medical and bio-mimetic engineering, i.e. the fields of artificial biology that aim to produce fully working bodies equipped with a sensorimotor system dynamically coupled with the environment: all machines truly capable of developing intelligence must be implemented in such system; however, as the technological reproduction of these bodily functions was not even supposable at the end of '40s, Turing had to concentrate his interests to the formal architecture of the algorithms that could better approximate and facilitate it (cfr. Turing 1948, p. 9).

The pioneering solution that Turing envisages in order to develop real embodiment is still based on a mechanistic representation of life and intelligence, but groundbreaking if seen as a developmental prototype: machine learning through embodied exploration of the environment, both free and under human supervision. The key model is a "child-machine" (Turing 1950, p. 456) that learns from attempting various acts of manipulation and circumspection, monitoring their results, and broadening the repertoire of available actions and desired goals. This model is inspired by the evolutionary adaptation of biological systems to their environmental niche: in fact, the "structure of the child machine" (assigned by the engineer and built-in), the "changes of the child machine" (dynamically developed through interaction), and the "judgment of the experimenter" (crucial to select and reinforce the successful conducts of the child-machine), correspond respectively–in Turing's proposal – to the "hereditary material", the "mutation", and the "natural selection" that determine the evolution of a real human infant (idem). Therefore, the living body is not only and not necessarily the result of natural evolution, as its characterizing function is to be the crucial mediator to learn from casual experience during development, i.e. (in Turing's terms) to dynamically update the contents of the table of behavior of a machine (here also called "the book of rules", cf. pp. 437–438) by fully exerting its very explorative/adaptive behavior. And this is a function that can be studied in a laboratory and possibly reproduced by artificial means.

Strikingly, this is the idea that today successfully inspires iCub, a tetrapodic anthropomorphic machine fully equipped with infant-like perceptual, motoric, and interactional bodily features, which in turn scaffold its capability to develop more and more complex action schemata (Metta et al. 2008; Cangelosi and Schlesinger 2015). The constant sensorimotor refinement of iCub's manual expertise, empowered by an advanced grasping control and force control systems, enables it to improve and indefinitely broaden its skills of hand-eye coordination, tool use, and fine object manipulation. Manipulative and explorative attempts, in turn, are used by iCub to perfect its recognition capabilities (actions, affordances, objects, and faces), and to update its routines to comprehend the causes and predict the effects of its physical interactions, also in conjunction with verbal commands, responsive gazing behavior, and elementary capabilities of joint attention.

It is probably while Turing was caressing the dream to father a similar artificial child that he started realizing that the body is the fundamental constraint/organizer/regulator in all the evolutionary and developmental processes of adaptation. However, as stressed by Cappuccio (2006) and Lassègue and Longo (2012), Turing's dream is still irremediably in debt with the metaphor of writing, a metaphor that once again he uses in these pages to describe the essence of the computational processes of the child-machine: "Presumably", says Turing, "the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.)" The persistence of this analogy (intelligence as a mechanical transfiguration of a writing system operating with contentful symbols) is precisely what brought him to misunderstand the role of the body in cognition.

28.5 Child-Machines and Multiple Realizability

Sixty years before the creation of iCub, Turing had already foreseen the importance of bodily functions for the development of a child-machine. Describing the child-machine's education, he regrets that "it will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it. It must be given some tuition." (Turing 1950, p. 456) Even if his tone seems humorous, Turing is seriously concerned that not having a body and a satisfactory social life might ultimately undermine the educational path of the child-machine project.

In spite of his farseeing vision, he surprisingly tended to deny their unique contingent causal role. Eventually, he decides to postpone the final struggle with this problem, trusting that one day engineering sciences will find a way to surrogate the relevant functional role of the body parts that the child-machine is not (yet) entitled to possess. "We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other" (idem). Turing's reference to Helen Keller is both inspiring and puzzling, from the perspective of embodied cognition. Helen Keller was famously born deaf-blind, and it is only thanks to an intensive educative program and constant assistance throughout her early childhood that she was able to overcome her incapacitating isolation, get a university degree, and eventually become successful as a writer and political activist. Turing brings this example to argue that even a serious lack of sensorial and communicative capability can be overcome by a functionally equivalent information processing apparatus (i.e., reducing real communication to exchanges of strings of written symbols from/to the isolated mind of Miss Keller). What he fatally overlooks is that the educational process that led Miss Keller to comprehend her instructor, and thus learn the first basic set of deaf-language signs, was actually intensely physical, at least in the beginning. This process was scaffolded by the participatory social modulation of affective and sensorimotor stimuli through informal hand gestures, guided by tactile feedback and motion. What is crucial in this methodology is precisely the bodily engagement of the child with the instructor (Ford 2011). In the quoted paragraph, it seems that Turing misunderstands the role of the body, depicting it as a set of modular and interchangeable peripherals whose function is merely instrumental and supplementary to the algorithmic work of a centralized computational agency.

However, if we look back at his 1948 work, we discover that at a certain stage he was ready to bet on a rather different cognitive role for the body. Foreshadowing the idea of a child-machine, in this paper, Turing had clarified even better that, in order to learn, a machine should implement a body capable of reproducing the entirety of the human organs and a dynamic neuronal network that is "unorganized", i.e. not regulated by pre-given sets of rules, but ready to progressively "auto-organize"

while learning from its experience of actual interaction with the world. "The machine should have a chance of finding things out for itself" (Turing 1948, p. 9) and this means "it should be allowed to roam the countryside" and, so to speak, make its own experience in the external world: by this reason, it should be endowed with legs and eyes (Turing also notes, with a sarcastic remark that probably targets the repressive conformism of his society, that "the danger to the ordinary citizen would be serious"). "However", complains the mathematician, "even when the facilities mentioned above were provided, the creature would still have no contact with food, sex, sport and many other things of interest to the human being". Interestingly, Turing lists among the pedagogical desiderata a series of bodily stimuli that humans typically consider enjoyable, as he considers them decisive to set the pre-categorial parameters that define desirable goals and normative conditions for the machine's dispositions to act. In operative terms, this is the primitive system of rewards that the machine's instructor can use to condition the machine's behavior. Indeed, Turing explicitly attributes a notable importance to pain/pleasure reinforcement mechanisms to guide the machine's learning experience and evaluate the success of its attempted actions, as these are the stimuli that the machine itself can use to measure its own behavior's distance from an competent performance. For the happiness of a behaviorist, to train a machine sensible to these stimuli would be more like conditioning an animal than compiling a computer program. But it is not indispensable to be a behaviorist to appreciate the phenomenological primacy of these intentional bodily experiences in scaffolding the development of intelligence.

Today we are inclined to believe that the conditioning force of bodily affects such as pleasure and pain is strictly inherent to the history of body-environment negotiations enacted through our experience (Bower and Gallagher 2013). This ecological view emerged only partially in Turing's conception (as highlighted by Wells 2005), which is still struggling to overcome the resilient belief that intelligence is envatted, i.e. reducible to input signals passively received by some sort of solipsistic Cartesian central module. In his conclusive remarks, Turing does not get to the point of declaring the body as a stringently (a priori) requisite for all kinds of intelligence, but admits that embodiment is an authentic and reliable strategy to actually reproduce intelligence, and moreover it is the most reliable one. In fact: "this method [unprincipled learning through spontaneously reinforced embodied interactions] is probably the 'sure' way of producing a thinking machine", although – adds Turing – "it seems to be altogether too slow and impracticable", probably also due to the technological limitations of his era. I think these observations prove that Turing never overlooked how real life contexts can directly incite action in actual bodies even when we do not have any internal model of those situations.

If we look at Turing's speculation on AI in the light of his work in theoretical biology we realize that he knew that actual bodies are necessary preconditions to make sense of real life contexts. According to the embodied approaches to cognition, if humans and other animals are able to adapt to the always exceeding complexity of real life circumstances, this can only happen through the negotiation offered both passively and actively by the body, through the details of physical

constitution, the retentive/protentive structure of its temporally extended immanent presence, and its embedded adaptive dispositions and skills. Turing himself was probably close to make a similar conclusion, against an instructionist view of biological and cognitive functions: his morphogenetic models can account precisely for the way the infinitely variable complexity of the contextual circumstances modulates the fine-grained constitution of the bodily functions even at the molecular level, adjusting them to the stochastic fluctuations of the environmental conditions. According to Longo, these "composite systems of action, reaction, and diffusion" allow a form of "calculus" that "entirely reside in the continuous dynamics of the forms. It is not anymore about an immaterial software, but constantly evolving unpredictable plasticity and materiality." (Longo 2009) In other words, the mechanical interactions between the material components of the system are not guided by some superimposed immaterial mental content; it is the very capability of the system to organize itself that provides it with intelligent adaptivity and flexibility. There is no meaning in life other than life itself, and this is what guides biological entities to develop thoroughly intelligent dispositions.

28.6 Multiple Realizability and the Frame Problem

However close to making this conclusion Turing was, he actually failed to spell it out – as he was clearly concerned about how "slow and impracticable" the embodied approach to AI could have been. The surrogative solution eventually proposed by Turing in 1948 (p. 9) to overcome the technological impossibility to endow a machine with a living body is "to try and see what can be done with a 'brain' which is more or less without a body, providing at most, organs of sight, speech and hearing". Turing has clear in his mind that the disembodiment of cognition can, at best, produce some formalistic approximation, or simplified prototype, of an intelligence dedicated to pre-regulated synthetic environments. Otherwise, he wouldn't have stressed repeatedly, and with such emphasis, the importance of implementing the right bodily features. It is true, however, that eventually Turing does not overcome the idea that the body is merely an extension, a replaceable appendicle, of cognition, rather than its inseparable backdrop and its condition of possibility.

After more than sixty years spent "trying and seeing" to build a body-less "electronic brain", most AI theorists abandoned this idea, turning themselves to the creation of intelligent functions that are deeply integrated, directly modulated, and actively scaffolded by their material constitution, deemphasizing more and more the role of encapsulated decision-making algorithms and centralized control systems: for example, the "soft" robotics (Pfeifer et al. 2012) and the bio-mimetics (Vincent 2009) approaches to AI, privilege solution in which the material background of cognition is not just a realizer of its logico-causal configurations, but its defining component. A theoretically relevant aspect of these trends is that they deemphasize the principle of multiple realizability (Putnam 1967a), which in philosophy is

typically associated to machine functionalism (Putnam 1967b) and a certain version of supervenience theory, and often exemplified by the distinction between the computational, the algorithmic, and the implementational levels of a cognitive architecture (Marr 1981): a cognitive process, exactly like a computer software, is multiply realizable when it depends on, but doesn't functionally identify with, the material details of its contingent implementation, which can therefore be realized by different configurations of "hardware" components (transistors, microchips, neurons . . .) as long as their causal relations bring about the same algorithmic and logical operations. Turing's proposal to substitute the living experience allowed through legs and eyes with equivalent sets of symbolic information manipulated through centralized algorithms clearly suggests that he was captivated by some version of the principle of multiple realizability, one that is at odds with the very idea of embodiment. In truth, at least some interpretations of multiple realizability (e.g., Wheeler 2010) don't exclude embodiment, and actually aim at accounting for the fine-grained details of the causal co-involvement of body and world in shaping cognitive functions at a sub-declarative level. That is why I believe that Turing's oversight is imputable to a particular conception of the body that is possibly derivable from the multiple realizability principle, but that is not implied by it.

This conception, which has often been the main philosophical alibi for the disembodiment of intelligence, is the one that postulates that the dichotomy of contents (the meaningful information represented by the system and manipulated by its functions) and vehicles (the meaningless realizers of the representational function) is a distinctive mark of cognition. In fact, this view, which is still mainstream in the cognitive sciences, maintains that the content/vehicle distinction is necessary in any explanatory discourse on mental functions because it assumes that the specific normativity of the mental can only be accounted for in terms of contents. Today, some anti-representionist trends in embodied/enactive cognition reject this view: i.e., the Radically Enactive Cognition, or REC (Hutto and Myin 2013). REC stresses that in no way this distinction could account for the basic structure of cognitive processes and that it actually is a categorical mistake to describe the functional realizers of cognitive processes as material containers of intangible informational contents, because any attribution of intentionality (mental states) to physical (biological or mechanical) processes inevitably violates the naturalistic assumptions behind the scientific study of the mind. Also other arguments, relevant to our discussion of Turing's legacy, could be used undermine the belief that detached representational content is a necessary component of cognition.

For example, the arguments related to the "frame problem", a philosophical term used to describe an artificial system's incapability to sense contextual relevance (Dennett 1987; Dreyfus 1992; Gallagher 2012; Cappuccio and Wheeler 2012). Dreyfus imputes this deficiency to the fact that, ignoring how any truly intelligent decision occurs against a bottomless background of preconditions, classical AI still relies on self-contained, limited representations or heuristic models of the world: but reality transcends any attempt to capture the complexity of a situation through its representations, i.e. content-bearing models or sets of instructions. In Turing machines, these representations are the symbolic models stored in the

tables of behavior of the machine. With regard to Turing's theoretical approach to computation, the frame problem is exactly the incapability of a discrete-state system (which uses finite sets of representations/instructions) to make sense of a transcendent continuous reality and to respond to its intrinsic relevance for a certain task.

Turing is aware of the existence of a similar problem, that he somehow addresses while discussing the "Argument of Informality of Behavior" against mechanical intelligence (1950, p. 452). This argument claims that a machine will never be able to reproduce the complexity of human behavior because "it is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances". Turing confesses to "agree" with the skepticism towards the idea of reducing intelligence to a rule-based system of decision making procedures, and since he also agrees that discrete-state machines are governed by rules, he finally admits that the human mind is not a machine of a logico-symbolic kind: this is sufficient to definitively erase Turing's name from the list of the alleged supporters of classical cognitivism. However, his overall judgment is nuanced. In fact, he denies that all human cognitive functions are irreproducible by mechanical means, and that all rule-based procedures should be excluded from consideration: Turing proposes a neat distinction between "rules of behaviors", which define the normative procedures used in explicit deliberation, and "laws of conduct", which describe the dispositions that prepare a cognizer to pre-reflectively respond to real-life complex circumstances. Conceptually, this distinction matches Dreyfus' separation between representational intelligence (mediated by contentful heuristics) and skillful coping capability (enacted through embodiment by contentless dispositions to real relevance). Therefore, one could well figure that Turing had already intuited that true intelligence could only be achieved by abandoning a rule-based representational architecture to avoid the Frame problem. At the same time, because the concept of machine that Turing had in mind was different than a deterministic step-by-step apparatus, he is adamant in asserting that even Frame problem-avoiding minds could be effectively imitated by machines of some type, though, probably, not of the discrete, logico-symbolic type: "we believe that it is not only true that being regulated by laws of behavior implies being some sort of machine (though not necessarily a discrete-state machine), but that conversely being such a machine implies being regulated by such laws".

28.7 Conclusions: The Frame Problem and the Symbol-Grounding Problem

The key issue that Turing never addresses is that the Frame problem should *necessarily* arise in discrete-state machines if we assume that the symbols used to instruct their tables of behavior form a system of contentful representations using principles of logical isomorphism/projection to capture the relevant aspects of a typical situation. The deeper problem then is how these representations could be

meaningful at all, as we don't have any general account of how representations could acquire their contextual content or be coupled to the practical background of real life.

This question discloses another problem that affects all the paradigms that rely on internally stored informational contents to explain intelligence, regardless that these paradigms assumed the transparency (symbolic approach) or the opacity (connectionist approach) of content: the "Symbol-grounding problem" (Harnad 1990). As illustrated by the 'the Chinese room' thought experiment (Searle 1980), this problem concerns the difficulty of explaining the origin of the meaning of the decisions taken by an intelligent system, provided that the representations manipulated to take decisions are conventional in nature, and therefore not meaningful at all to the system itself. The problem of the non-derivability of meaning (and thus relevance) from mere syntactical procedures, is even better illustrated by the historical circumstances in which understanding the meaning of an encoded message requires, first of all, to know the broader real-life context and the underlying communicative intention, rather than blindly apply a set of algorithms for symbolic substitution. Turing probably had to intensely struggle with the intrinsic semantic emptiness of conventional symbols, during his code breaking experience at Bletchley, which certainly solicited him to appreciate the collaboration with linguists, anthropologists, and experts of German culture, beside logicians and mathematicians: the peculiar participatory experience of trying to make sense of an encrypted message probably made him aware that the pragmatic meaning of a string of symbols does hardly come only from their syntactic rules of combinations, as it first of all emerges from the broader context (historical, cultural, and bodily) of their actual use. Symbols are never sufficient to recreate their context of meaning, if the agents who manipulate them are not already embodied in it: only a staunch formalist mathematician would not recognize it. Turing was not a formalist of this kind but, if he occasionally failed to remember this lesson, in his late speculation on AI, this is because – still influenced by the heritage of a formalist thought – he was never entirely capably to get over the belief that semantic content is a commodity that could be conventionally attached to symbols and freely transferred from a string of symbols to another.

Summing up, in line with the anti-representationist approaches to basic cognition like REC, I believe that if Turing eventually failed to fully value the cognitive-developmental role played by the body, this was not because he proposed a computational and functionalist model of the mind, but because he tacitly assumed the content/vehicle dichotomy as a primitive of that model: in 1950, he still believes that intelligence is a matter of accessing contents, allegedly detachable and transmittable from a sender to a receiver through a string of conventional symbols. However, because Turing revolutionary approach to the mechanization of symbolic processes was not aware of this implicit metaphysic, his belief was not dogmatic. On the contrary, as the British mathematician was keen on valuing the situated context of real intelligence, in his work we can find certain premises to think cognition as embodied and represention-less. However, as we have seen, Turing's approach to this issue is uncertain: on the one hand, its computational machines promoted

a tacit symbolization (representational and formal inscription) of the mechanisms of our minds, suggesting that this functioning is content-bearing in nature; on the other hand, his proposal of child-machines provided us with important reflections to abandon the representational paradygm and rethink the meaning of the grammatical mataphor.

References

Boden, M. A. (2006). *Mind as machine: A history of cognitive science* (Vol. 2). Oxford: Oxford University Press.

Bower, M., & Gallagher, S. (2013). Bodily affects as prenoetic elements in enactive perception. *Phenomenology and Mind, 4*(1), 78–93.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159.

Cangelosi, A., & Schlesinger, M. (2015). Developmental robotics. From babies to robots. Cambridge, MA: MIT Press.

Cappuccio, M. (2005). Un'ipotesi controfattuale: la macchina di Turing ideografica. In L'eredità di Alan Turing. 50 anni di intelligenza artificiale (pp. 231–244). Milano: Alboversorio.

Cappuccio, M. (2006). *Alan Turing: l'uomo, la macchina, l'enigma. Per una genealogia dell'incomputabile*. Milano: Alboversorio.

Cappuccio, M., & Wheeler, M. (2012). Ground-level intelligence: Action-oriented representation and the dynamic of the background. In Z. Radman (Ed.), *Knowing without thinking*. London: Palgrave Macmillan.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics, 58*(2), 345–363.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.

De Jaegher, H., Di Paolo, E., & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences, 14*(10), 441–447.

Dennett, D. (1987). Cognitive wheels: The frame problem in artificial intelligence. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence*. Norwood: Ablex.

Dreyfus, H. L. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.

Dreyfus, H. L. (2002). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences, 1*, 367–383.

Dreyfus, H. L. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In P. Husbands, O. Holland, & M. Wheeler (Eds.), *The mechanical mind in history*. Cambridge, MA: MIT Press.

Ford, J. (2011). Helen Keller was never in a Chinese room. *Minds and Machines*, *21*(*1*), 57–72.

Freeman, W. J. (2000). Neurodynamics: An exploration in mesoscopic brain dynamics (Perspectives in neural computing).

Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Oxford University Press.

Gallagher, S. (2009). The key to the Chinese room. In K. Leidlmair (Ed.), *After cognitivism* (pp. 87–96). Dordrecht: Springer.

Gallagher, S. (2012). Social cognition, the Chinese room, and the robot replies. In Z. Radman (Ed.), *Knowing without thinking*. London: Palgrave Macmillan.

Gallese, V. (2009). Mirror neurons, embodied simulation, and the neural basis of social identification. *Psychoanalytic Dialogues, 19*, 519–536.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42*, 335–346.

Haugeland, J. (1978). The nature and plausibility of cognitivism. *Behavioral and Brain Sciences, 1*, 215–260.
Hendriks-Jansen, H. (1996). *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought*. Cambridge, MA: MIT Press.
Herrenschmidt, C. (2007). *Les trois écritures. Langue, nombre, code*. Paris: Gallimard.
Hobson, P. (2002). *The cradle of thought*. Oxford: Oxford University Press.
Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without contents*. Cambridge, MA: MIT Press.
Kleene, S. C. (1936). Lambda definability and recursiveness. *Duke Mathematical Journal, 2*, 340–353.
Kleene, S. (1952). Introduction to metamathematics. Amsterdam: North-Holland.
Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.
Lassègue, J. (1998). *Turing*. Paris: Les Belles Letters.
Lassègue, J., & Longo, G. (2012). What is Turing's comparison between mechanism and writing worth? In S. B. Cooper, A. Dawar, & B. Loewe (Eds.), *Computability in Europe* (pp. 451–462). Berlin/Heidelberg: Springer.
Longo, G. (1999). The difference between clocks and Turing machines. In A. Carsetti (Ed.), *Functional models of cognition* (pp. 211–232). Dordrecht/Boston: Kluwer.
Longo, G. (2002). The constructed objectivity of mathematics and the cognitive subject. In M. Mugur Schacter (Ed.), *Proposals in epistemology. On quantum mechanics, mathematics and cognition* (pp. 433–463). Dordrecht/Boston: Kluwer.
Longo, G. (2008). Laplace, Turing and the "imitation game" impossible geometry: Randomness, determinism and programs in Turing's test. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing test* (pp. 377–413). Dordrecht: Springer.
Longo, G. (2009). From exact sciences to life phenomena: Following Schrödinger and Turing on programs, life and causality. *Information and Computation, 207*, 545–558.
Marr, D. (1981). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: The MIT Press.
McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.
Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition, PerMIS '08. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems* (pp. 50–56).
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.
Minsky, M. (1974). A framework for representing knowledge. MIT Lab Memo # 306.
Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
Noe, A., & O'Regan, J. K. (2002). On the brain-basis of visual consciousness: A sensorimotor account. In A. Noe & E. Thompson (Eds.), *Vision and mind: Selected readings in the philosophy of perception*. Cambridge, MA: MIT Press.
Nolfi, S., & Floreano, D. (2000). *Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines*. Cambridge, MA: MIT Press.
Petitot, J., Varela, F., Pachoud, B., & Roy, J. M. (Eds.). (1999). *Naturalizing phenomenology: Contemporary issues in phenomenology and cognitive science*. Stanford: Stanford University Press.
Pfeifer, R., Lungarella, M., & Lida, F. (2012). The challenges ahead for bio-inspired 'soft' robotics. *Communications of the ACM, 55*(11), 76.
Pinker, S. (2002). *The blank slate*. New York: Penguin.
Putnam, H. (1967a). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
Putnam, H. (1967b). The nature of mental states. In D. M. Rosenthal (Ed.), *The nature of mind* (pp. 197–203). New York: Oxford University Press.
Reddy, V. (2008). *How infants know minds*. Cambridge, MA: Harvard University Press.

Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Volume 1: Foundations). MIT Press: Cambridge, MA.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417–457.

Shapiro, L. (2011). *Embodied cognition*. New York: Routledge.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem: A correction. *Proceedings of the London Mathematical Society, 43*, 544–546.

Turing, A. M. (1948). Intelligent machinery. Report for national physical laboratory. In: Collected works (Vol. 1).

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 50*, 433–460.

Turing, A. M. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London, Series B, 237*, 37–72.

van Gelder, T. (1991). Classical questions, radical answers: Connectionism and the structure of mental representations. In T. Horgan (Ed.), *Connectionism and the philosophy of mind*. Dordrecht/Boston: Kluwer Academic Publishers.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: The MIT Press.

Vincent, J. F. V. (2009). Biomimetics, a review. Journal of engineering in medicine. *Proceedings of the Institution of Mechanical Engineers. Part H, 223*(8), 919–939.

Wells, A. (2005). *Rethinking cognitive computation: Turing and the science of the mind*. London: Palgrave Macmillan.

Weyl, H. (1985). Axiomatic versus constructive procedures in mathematics. In T. Tonietti (Ed.), *The mathematical intelligence* (7)4. New York: Springer-Verlag.

Wheeler, M. (2010). In defence of extended functionalism. In R. Menary (Ed.), *The extended mind*. Cambridge, MA: MIT Press.

Wittgenstein, L. (1980). *Remarks on the philosophy of psychology* (Vol. 1). Oxford: Blackwell.

Chapter 29
Heideggerian AI and the Being of Robots

Carlos Herrera and Ricardo Sanz

Abstract Current Heideggerian AI (HAI) is the attempt to revise the fundamentals of Artificial Intelligence based on Heidegger's philosophy. While the debate is much monopolized with questions regarding the role of representations, there is overall agreement that HAI should be conceived to foster development of AI techniques, on the assumption that Heidegger's ontological analysis of humans (Dasein) should apply to artificial systems. We argue this is inconsistent with Heidegger's philosophy, as it denies ontological meaning to categories such as robot and human, considered the same type of beings. The aim of this paper is to steer HAI towards the question of our pre-ontological notions of artificial systems, and robots in particular. We present a provisional ontological analysis that considers robots specific, non-human and non-animal beings, which we derive from the relationship between robots and work. Robots are those machines that perform human labour – because in practice they can only transform it, their being is one that cannot be fulfilled.

Keywords Martin Heidegger • Heideggerian AI • Hubert Dreyfus • Ontology • Robotics

29.1 Introduction

Martin Heidegger is widely considered one of the central intellectual figures of the twentieth century, with his influence spreading throughout philosophy, arts and the social sciences. He may even be considered a central piece in the shift of attention in the cognitive sciences from detached thought to embodiment and situatedness. For instance, Andy Clark's book "Being There" (Clark 1998) clearly signals affinity with Heidegger's thesis that body and world are both constitutive of the human being, Dasein (often translated to English as Being-there).

C. Herrera (✉) • R. Sanz
Universidad Politécnica de Madrid, Madrid, Spain
e-mail: carl.her@gmail.com; Ricardo.Sanz@upm.es

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_29

Heideggerian thought also played an important role in the change in paradigm that Artificial Intelligence (AI) has experienced over the last few decades. The first critiques to traditional AI, developed by Hubert Dreyfus, called for a reappraisal of one of the fundamental techniques used for the design of intelligent machines – that of symbolic representation (Dreyfus 1972). Heidegger stressed how human everyday coping does not rely on rational and inferential thought processes, which suggests that mental representations need not be the universal mechanism for intelligence.

Phenomenological alternatives to traditional assumptions, such as the significance of body and situation, have in turn become buzzwords of the new Situated and Embodied AI paradigm (Brooks and Steels 1995). Its goal is no more to create super intelligent information-processing computers, but skilful robots. Within this paradigm, Heideggerian AI (HAI) has been coined as the subfield dedicated to investigating the different ways in which Heidegger's philosophy may inform research in AI. Even though the boundaries between *Situated and Embodied AI* and HAI are rather thin, at least one difference should be found: HAI aims to be consistent with Heidegger's philosophy. This consistency is our main concern here.

The dominant approach led by Dreyfus considers Heidegger's analysis of everyday human action, as found in Being and Time, and contrasts its phenomenology with that of intelligence as captured by rule-based artificial systems. On the basis of the limitations of such systems he argued that, in order for artificial systems to progress, they must be phenomenologically comparable to humans. This approach has been fruitful both for the popularisation of Heidegger in this area and the advancement of embodied-situated AI – and remains an important open area in AI and robotics research.

How to interpret Heidegger in relation to the design of artificial intelligent systems is of course a matter of speculation, as he did not discuss the issue directly, and his philosophy is complex enough to sustain different interpretations. Even within the Heideggerian AI background, Dreyfus denounces "Wheeler's cognitivist misreading of Heidegger" (Dreyfus 2007), regarding the role representations should play in artificial systems. There is nevertheless overall agreement on the needs "for a positive account of Heideggerian AI and of an underlying Heideggerian neuroscience" (Dreyfus 2007). But, can this by itself be considered congruent with Heidegger's philosophy?

It would appear not to be if we take into consideration Heidegger views on science, technology and their actual significance for us. One of his concerns is that technology does not necessarily make humans more free, because it can be seen to demand humans to serve technology as much as the other way round. Some have even suggested that the AI project would have probably appeared as the exemplar of what is wrong with technology for Heidegger (cf. Masís 2009[1]).

[1] "¿No comprendería Heidegger, en efecto, el proyecto de la IA como el fastigio de la metafísica, como la tecnificación más nefanda?" Masís 2009 (Wouldn't Heidegger understand the AI Project as the pinnacle of metaphysics, as the most abominable technification (own translation).

This paper argues that a genuine Heideggerian perspective on AI cannot be found either in the faithful submission to AI goals and aspirations, nor to the complete disregard of this form of technology. The Heideggerian perspective on AI has to be developed as a way of questioning akin to that exercised by Heidegger. For this, we take into account three of Heidegger's central themes: the question of being, the relationship between science and philosophy, and the essence of technology.

We first review the development of Heideggerian AI as we know it, from Dreyfus initial critique to further developments in the field. We then discuss Heidegger's notion of being, technology and science, and their relevance to develop a Heideggerian AI that fits into this background. We then reappraise the issue of Heideggerian AI as the question on the being of robots, and offer a preliminary discussion that should serve as a first step in the development of the area.

29.2 A Brief History of Heideggerian AI

Hubert Dreyfus is undoubtedly the founder and main defender of Heideggerian AI. If we pay attention to how Dreyfus' work developed over the years, we can distinguish a major shift, not simply on Dreyfus position, but also in AI. The "first Dreyfus" was thought and perceived by the AI community as an outsider, a cross-disciplinary academic, attacking an alien discipline such as AI (e.g. Papert 1968; Buchanan 1972; cf. McCorduck 1979). Several decades later, nevertheless, his work is seen as inter-disciplinary, providing a positive contribution to AI.

In the early 60s the world of psychology and incepting cognitive science was dominated by the computer metaphor of the mind. This metaphor provided a twofold assumption. First, that computers have the necessary means to perform any intelligent operation available to humans; second, that intelligence must be understood as formal operations of the mind.

The second assumption has been commonplace in philosophy since Descartes within the philosophical school that may be called *rationalism* (cf. Taylor).[2] Though not always acknowledged, classical themes in rationalistic philosophy re-emerged in the form of common sense assumptions in the AI project (cf. Kenaw 2008). When Dreyfus, a researcher with a background in phenomenology and continental philosophy, become acquainted with the AI research program, he was struck by the naivety of some AI researchers who claimed to finally have overcome philosophy, when in fact they were simply formalizing the assumptions of the rationalistic tradition.

[2]The dominant view of rationalism "offers us a picture of an agent who in perceiving the world takes in 'bits' of information from his or her surroundings and then 'processes' them in some fashion, in order to emerge with 'this picture' of the world he or she has; who then acts on the basis of this picture to fulfill his or her goals, through a 'calculus' of means and ends" (Taylor 1993, p. 204).

After centuries of domination in western philosophy, rationalism had been seriously challenged, in special by the work of Heidegger (Taylor 1995, p. 61). Dreyfus had the intuition that if Heidegger's work had been effective in overcoming rationalism, it could also apply to the somehow childish AI program. He then looked for ways in which a critique to AI could be articulated. Researchers excess of optimism, claiming to have put an end to philosophy and to be close to developing machines that would imminently match human intelligence (e.g. Newel and Simon 1963) made the task easier.

The first attempt was developed by Dreyfus in a report entitled "Alchemy and AI" commissioned by the RAND corporation (Dreyfus 1965). Despite the influence of phenomenology, in this paper Dreyfus attempts to articulate philosophical ideas in a language that would be understandable and useful for AI researchers, dealing with issues that arise from the research goals and methods of AI. He successfully identified some of the core problems of classic AI without resorting to Heidegger. Despite this effort to talk engineering, "Alchemy and AI" received a hostile reception and was widely perceived as an attempt to demolish the field (Cf. McCorduck 1979). There was certainly an attack on the blindness of the community towards some of its failures and limitations – but as a consequence of the paper, such blindness became exacerbated in a complete dismissal of Dreyfus well-articulated criticism.

Dreyfus argued that the internal management of a indeterminate number of representations could never be sufficient for the development of skilful activity. The problem is not to posses the information, but to deploy the relevant knowledge in the right situation. Humans achieve this, not through processing a large database of detached knowledge, but through being coupled to our tools and systems. A world of significance is laid before us, so that representations are only required once our coupling breaks down.

Dreyfus critique of AI resonated with a number of AI researchers such as Agre, Winograd or Flores (Winograd and Flores 1986; Agre 1997). These researchers attempted to move AI from the detached perspective to a situated one, for instance investigating the design of agents that do not rely on context-free symbolic representations or internal model-based planning (cf. de Beistegui 2005).

The critique of representational AI principles only hit mainstream AI when developed by Rodney Brooks at the MIT (Brooks 1991). Essential for the success of this critique was the consideration of robotics as a partly independent challenge form previous AI work on formalising abstract intelligence. For Brooks, the traditional AI approach to robotics had great difficulties to produce operational systems, because instead of considering the problem of embodied interaction, they diverted their attention to formalising abstract thought. Brooks challenged the notion of representation as the basic unit of abstraction for developing intelligence, which has been a recurrent theme in Heideggerian AI. Behaviour-based approaches, such as Brooks' subsumption architecture, can avoid some of the problems of representational approaches, although it can by no means be considered sufficient for the development of AI.

By bringing robotics to the forefront, Brooks was not challenging the dreams of AI researchers, but their actual work, in a clear and apparently effective way. Brooks' thesis was followed by a rapid change in mood in the AI community, and a reorganisation of the field. In a short span of time, the once marginal and radical critique of the establishment became a common place. The huge successes that before were celebrated under the umbrella of AI now became computer science, and the challenge of creating intelligent artefacts was taken over by the field of robotics. Embodiment and situatedness became buzzwords, and although Heidegger's influence is still taken with caution, it is undoubtedly among the sources of new AI and cognitive science.

Despite this influence, the idea of "Heideggerian AI" as a research project has been advocated only by a small number of researchers, led by Dreyfus. For the rest, Heidegger is just one among other phenomenological sources (often less preferred than Husserl or Merleau-Ponty). Thus for the majority the question is not whether Heidegger's influence can be seen as a development of Heideggerian philosophy, but whether it provides support for their research goals. Yet for those who claim to being true to Heideggerian philosophy (as it happens with Dreyfus) we should ask whether the common interpretations of his philosophy are rightly so.

29.3 Clarifying the Role for Heideggerian AI

The interpretation of Heidegger in the consideration of artificial intelligence will necessarily be a subject of contention. Nevertheless, there is overall agreement in seeing Heideggerian AI as the attempt to reshape the fundamental notions of cognitive science, often incorporating Heideggerian terminology, with a commitment to the advancement of AI. In this section we discuss to what extent this view is congruent with Heidegger's philosophy.

We first draw attention (Sect. 3.1) to Heidegger's own explicit concerns regarding the task of bringing his philosophy into a particular scientific enterprise. Because Heideggerian AI is in fact such an attempt, we can consider his views essential to understanding the role of Heideggerian AI. We also discuss Heidegger views on technology and how they may relate to AI (Sect. 3.2).

29.3.1 AI, a Discipline in Crisis

Heidegger delivered 'Introduction to Philosophy' courses at the University of Friburg, attended by a great number of students from different disciplines (Heidegger 2001). He was concerned with letting them understand philosophy in relation to science. He talked about the crisis of the sciences (a popular theme in his time), somehow referring to parcels of discontent within contemporary sciences, such as physics or biology in his time, or cognitive science nowadays.

The road to autonomous intelligent machines is proving more difficult and challenging that it was first expected – not only because of a lack of instruments and methods, but because it is from the beginning unclear what is meant to be achieved. Researchers routinely express dissatisfaction with progress in the state of the art. Roboticists construct hypotheses, often based on neurological, psychological or philosophical theories, of what autonomy means and what mechanisms are required. Ad-hoc implementations of simple robots are presented to support the theories, but few conclusions can taken. Competing paradigms offer different avenues, but hundreds of experiments are not sufficient to decide between them.

For Heidegger, dissatisfaction with current bodies of knowledge and available methods gets reflected in attempts to redefine fundamental notions. Often, philosophy is brought-in to catalyse investigations about the foundation of a particular science (cf. Heidegger 2001). This was as true then as it is now, as we have seen in cognitive science.

Heidegger is nevertheless explicitly critical with trying to "reform a science with the help of Heideggerian terminology", an attempt which he describes as blind eagerness and agitation (Heidegger 2001, p. 52). The main reason is that the so-called crisis is not a malaise to be resolved once and for all through fundamental questions and research programs, but the driving force of such a science. For a science to exist the crisis needs to become alive and experienced, and not simply to contribute to its progress– but first and foremost need for such a science to exist.

This crisis calls the scientist to delimit fields of research in new ways, as science is a mode of enquiry based on a previous determination of the region of beings studied (see Sect. 4.1 – Heidegger's theory of science). Science crystallizes in a body of knowledge, but it cannot be reduced to a set of true propositions about the world. If we gathered all scientific knowledge and used it without challenging it, we would not be doing science anymore. Much more fundamental is the drive of researchers to rethink nature, to try to define and delimit it so that their questions can be answered unequivocally.

Under this perspective, the so-called paradigm change is no more than an expression of the crisis latent in the AI project. That AI may be a project in crisis or an immature science should not be considered from Heidegger's point of view as a critique – it is a sign that it is very much alive, that the intuition that artificial intelligence is possible still needs to be grasped in full. The search for the capacity to model human intelligence is so huge that we are far from the comfort zone of knowing how to ask the right questions. Heidegger, among other philosophers, have been used as catalysers for this crisis, but the role of Heideggerian AI is not to resolve the crisis, but to bring it to the forefront. To a certain extent this was achieved by Dreyfus's work and by the critical atmosphere that developed over the years. But the critique applies to the traditional representational approach to building robots. In the next sections we argue that the real issue is indeed not how robots are built, but what they are.

29.3.2 Concerning Technology

Before we move forward we will consider Heidegger's views on technology. These are for many (and with a reason) considered to take a moral stance against technology and the modern scientific world (Dieguez 2009). "Everywhere we remain unfree and chained to technology, whether we passionately affirm or deny it" (Heidegger 1982). Using Heidegger for the advancement of AI might sound to some hard-core Heideggerians as a contradiction in its own terms. At least, this is how early AI researchers interpreted Dreyfus initial critique – which nevertheless has made AI a central topic of his own research.

Heidegger critique of the technological mind is summarised in the notion of "enframing" (Heidegger 1982). Here he points to a simplified experience of the world due to a utilitarian framework. For technology, understood as the means to some ends, all beings in the universe may be regarded as potential resources. This applies not only to entities in nature, such as rivers, animals or fossil fuels, but also to humans, as technology also covers modes of organisation and control of humans. In this sense, Heidegger claims that all sciences will eventually be reduced to a more fundamental one that is cybernetics – where every system and process is understood as instrumental to some end (Heidegger 1993).

This cybernetic perspective implies that we are not interested in ontological questions, because all entities fall under the same umbrella. "The sciences are now taking over as their own task ... the ontologies of the various regions of being ... categories, which are allowed only a cybernetic function, but denied any ontological meaning" (Heidegger 1993). In a cybernetic, systems understanding of the world, the "entity" is always arbitrarily defined, that is, arbitrarily separated from its environment (Klir 1991). The system that is the object-of-interest is hence arbitrarily bounded. This implies that the "being of an entity" in the world is shaped by our epistemological decisions more than by an ontological analysis of the reality under study.

In other words, the technological world makes us forget the question of being, which is for Heidegger's of uttermost importance. To pursue technological development without ontological insight can make us plunge "into an all-pervasive fog in which we wrestle with we know not what" (Pattison 2000, p. 65). The role of philosophy must be to re-awaken question that traditional metaphysics have buried over the centuries (Heidegger 1962), in order to relate to things authentically. This critique could suggest that the role of Heidegger's philosophy is primarily to counteract the effects of technological enframing. Where technology makes us see the world in an instrumental (non-authentic) way, his philosophy offers a way to acknowledge ontological meaning as fundamental. Heideggerian AI would seem to go in the opposite direction to AI.

Nevertheless, for Heidegger the advancement of technology is a human necessity or destiny. It is not a question of whether we like it or not. We can assume that,

sooner or later, in one form or another, the development of autonomous systems will reach levels only dreamt of, and therefore now only poorly understood. And this will surely impinge on the way we experience the world, and on human existence generally. Here is where the essence of technology lies for Heidegger, and thus should be the primary concern for a Heideggerian perspective on AI. To underestimate the importance of AI, or to reduce it to the general analysis made of technology, is not Heideggerian, but contributes to the phenomenon of enframing.

29.4 Heideggerian AI and the Foundations of Cognitive Science

29.4.1 Heidegger's Theory of Science

AI is intertwined with contemporary cognitive science, and thus can be considered a scientific enterprise. As mentioned before, researchers (often sympathetic to Heidegger) have argued for a paradigm change in cognitive science of the sort discussed by Kuhn (2012). Yet, Heidegger's theory of science has been largely ignored in the Anglophone world in general (cf. Rouse 2008), and in Heideggerian AI in particular.

Heidegger distinguishes two kinds of enquiries: ontic and ontological. Ontic research is directed to facts relating to entities that belong to a well-defined domain, and thus it is comparable to the sciences. Ontological investigation, on the other hand, is directed to the being in question, the domains of Beings or the notion of Being itself.

For Heidegger "ontological inquiry is indeed more primordial, as over the ontical inquiring of the positive sciences (1962 p. 30)". The sciences are always 'founded' on some ontological pre-understanding that defines the area of study. "Basic concepts determine the way in which we get an understanding beforehand of the area of object-matter underlying all the objects a science takes as its theme, and all possible investigation is guided by this understanding" (1962, p. 30).

This 'theory of the sciences' is summarised in the introduction to Being and Time (Heidegger 1962, p. 29):

> Being is always the Being of an entity. The totality of entities can, in accordance with its various domains, become the field for laying bare and delimiting certain definite areas of subject-matter. These areas, on their part (for instance, history, Nature, space, life, Dasein, language, and the like), can serve as objects which corresponding scientific investigations may take as their respective themes. Scientific research accomplishes, roughly and naively, the demarcation and initial fixing of the areas of subject-matter. The basic structures of any such area have already been worked out after a fashion in our pre-scientific ways of experiencing and interpreting that domain of being in which the area of subject–matter is itself confined. The 'basic concepts' which thus arise remain our proximal clues for disclosing this area concretely for the first time. And although research may always lean

> towards this positive approach, its real progress comes not so much from collecting results and storing them away in 'manuals' as from inquiring into the ways in which each particular area is basically constituted ...

A scientific enterprise is always built on an ontological understanding of the region of nature it deals with. In simple words, for someone to do research in physics, chemistry or psychology, she must already have an idea of what these areas are. This pre-understanding is not a 'folk science', the set of pseudo-scientific beliefs a layman can have. More basically, it refers to our experience of the world that makes possible to become aware of the field in question.

Real movement in the sciences occurs not just through the development of methods and empirical findings, but when the limit of the region of beings it deals with goes through a revision. The typical example is that of physics, but it could also be drawn from modern psychology. The object of traditional psychology is the human psyche, and the conditions that may result in dysfunctions. Behaviourism considered behaviour as the fundamental object of research, therefore extending the field of psychology to animals. The revolution of the computer metaphor of the mind shifted the object of research to systems with the capacity of processing information in ways that can allow representations of the world. Although the paradigmatic object was still the human psyche, empirical research on computers was equally valid.

The recent shift in cognitive science results from another change in the delimitation of the field of study. Now the proper objects are embodied systems capable of sustainable interaction with their environment. This results in methodological changes in cognitive science. Human mental operations are interesting insofar they result from some form of embodied interaction. Robots, on the other hand, have replaced computer and information processing systems as a proper object of empirical research.

From our discussion of Heidegger's theory of science emerges a possible role of Heideggerian AI – as the attempt to set the grounds of a new science of cognition, by helping do what science requires to advance, that is, to determine a field of research. In this line, it has been argued that Heidegger's philosophy "provides a conceptual framework for the cognitive science of today and of the future" (Kiverstein and Wheeler 2012).

29.4.2 Robotics as a Distinct Region of Beings

Fundamental questioning can help determine the region of beings studied by a science. For both cognitivist and embodied cognitive science, the object of study are adaptive processes found not just in humans, but also animals and artificial systems. The grouping of Dasein with other beings under the same science is entirely consistent with Heidegger's views. In fact, the same happens in biology (covering humans as much as microorganisms) or even in physics (which covers all physical entities). Ontological difference, which is for Heidegger of uttermost importance,

becomes nevertheless blurred when we consider a single region of beings.[3] There is no contradiction if our analysis remains at the ontic level – features of cognitive-behavioural processes that humans may share with animals and machines, common ontic abstractions. This is different to talk about the form of being associated to such potential identity – the ontological analysis of Dasein needs not apply to beings that are non-human.

AI has traditionally the conceptual framework of cognitive science. The identification of robots with animals and humans sets AI with a clear target, a roadmap which if achieved would be a great success. AI Researchers, including those that advocate for a Heideggerian approach, have adopted this roadmap. Heideggerian AI, rather than identifying ontological differences, takes Dasein's ontological structure as the blueprint of robot agency, and little is discussed about the special being of robots. The assumption is that the goal of AI should be to close the ontological gap, create a faithful replication of human intelligence. "If we can't make our brain model responsive to the significance in the environment as it shows up specifically for human beings, the project of developing an embedded and embodied Heideggerian AI can't get off the ground ... we can, however, make some progress towards animal AI" (Dreyfus 2007).

Heideggerian AI has thus been concerned with how robots should be constructed (e.g. concerning the use of internal representations), and whether that would allow robots to have an ontological structure like Dasein. But the mismatch is necessary: while on the one hand we consider the ontic structure of robots, on the other we have the ontological analysis of humans. In other words, even if the problem regarding internal representations were avoided, there would still be an unsurpassable gap between ontic descriptions of robot technology and an ontological analysis of Dasein.

Heideggerian AI should first question our pre-ontological notion of robots. This is still tied-up with that of animal and human, in science and fiction. Only a few decades ago our concept of an ideal robot was that of a clumsy scientist with vast databases of knowledge and infallible capacity for rational argument (think for instance of C3PO in Star Wars). In the last 20 years, we are prone to think of extremely agile insect-like machines (e.g. the robots in The Matrix). This change reflects and is reflected in actual AI. Toda (1962), Dennett (1978) and Brooks were among the first to suggest that what was required for a new development in robotics was a change in our concept of robot. They convinced much of the autonomous robotics community that no longer should the human mind be the blueprint for robotics, but the capacity to cope with the real world already found in insects. Robotics should follow the line of evolutionary complexity, and only target human intelligence whence animal intelligence is achieved (Brooks 1991).

[3] Are robots imperfect copies of biological systems, the "purest" form of autonomy? Or are animals and humans genuine examples of robotic systems, i.e., "biological robots"? The problem may not be so much the reductionist view of humans, but the dismissal of the possibilities of robotics beyond the artificial animal or human.

This desire to redefine the area of study of robotics even motivated the use of the term animat and artificial life (Langton 1986; Wilson 1991), in an attempt to get rid of the connotations of our pre-ontological notions of robot and artificial intelligence. This attempt, although permeates in much of current research, has nevertheless not been radical enough, even if some questions have arisen regarding the difference between biological and robotic autonomy (e.g. Ziemke 2008).

We suggest that Heideggerian AI should take a different turn, taking up the ontological analysis of robots, the central question being: What is a robot? This may sound like a simple, even gratuitous, question – but these are, in Heidegger's words, the most difficult to ask. Focussing on ontological characteristics of robots makes Heideggerian AI step back from the main objective of AI; to advance the techniques available for the development of robots. Instead, it must reconsider the being of robots and challenge some of the assumptions regarding robots that permeate society and scientific research – our pre-ontological understanding of robots, the primary way such beings are disclosed and the set of intuitions about what robots are.

Heideggerian AI should be concerned with a new ontological analysis of robots – considering robots specific, non-human and non-animal beings, ontologically distinct to other technological artefacts. In other words, the ontological gap between Dasein and machine should not be a worry for Heideggerian AI, but the first stone on which a further understanding of AI should be built. This can be instrumental to fundamental revisions in the area of robotics and its development, but its aim is not to derive a precise framework where a new science could develop. This is still the task of science and technology, and driven by experience and experimentation. In other words, the real test of Heideggerian AI is not whether it can produce positive results (as implied in Dreyfus 2007), but whether it could successfully thematise the being of robots.

It is not the role of Heideggerian AI to solve the puzzle for AI development, but to unveil the being of robots, now shadowed by the idea of "an artificial human". The more we uncritically identify robots with animals and humans, the further we are from becoming aware of what robots really are. An ontological analysis may find that the being of robots is somehow derivative of Dasein, but that cannot be a premise of the investigation. Of course, that technology could advance so far to make humans and autonomous robots the same kind of beings will remain a possibility. But our premonition is that the goal of AI is not to bridge this gap, but to allow the appearance of new machine-like ways of being in the world – while transforming Dasein.

29.5 Questioning the Being of Robots

29.5.1 Robot Workers

Although Heidegger's way of enquiring had been exercised in several different areas of being, there are no clear guidelines to how ontological analysis should be

conducted. The following discussion is our attempt to open up the question on the being of robots, to acknowledge that enquiring about the ontology of robots makes sense. Thus we can only offer an *incomplete and provisional analysis.*[4] Some of our insights will meet criticism and phenomenological objections, but our goal will be achieved if we can motivate more systematic analysis and the development of interest in the ontological analysis of robots.

When we begin to question the being of robots, we must ask what region of being we are dealing with, and what characterises their existence. We first find that robots form a heterogeneous region of beings: we must count operative industrial and service robots, and robots in research laboratories, often not fully operational but serving as proof-of-concept for potential robot technologies.

Researchers often "imagine" the impact that a concrete technology might have in future robots. They build robots, but the claims they make are the result of escalating their vision to future robots. We thus must take into account, together with factual robots, potential robots, the robots that are yet to be and that robotics research aims to disclose. We must even consider fictional robots, as science fiction is a strong source of our pre-ontological understanding of robots.

That robots form a heterogeneous region of beings demands that we identify something characteristic to all robots, to their way of being.[5] We know this has nothing to do with representations: there are robots that rely heavily on representations; there are some that do not at all. In other words, even if some design approaches may be more powerful and far-reaching than others – or successful where others fail - none can claim hegemony over the definition of robot. The defining characteristics of robots cannot be found in their design.

For Heidegger the central ontological feature of humans is being-in-the-world. The being of robots too is situated, although maybe in a different way to humans. It is not just that a robot must always be placed in a physical environment – it must also behave towards its environment in a way that reflects a network of significance – goals, concerns, meaningful events etc. The behaviour of a robot must make sense in its world – or rather, in our world. For robots are made for a reason, and normally involves carry out some task that is useful – they do a job. That is, they play a role in the organisation of labour, replacing and transforming human work, and thus reducing labour costs. This, we argue, is an ontological feature of robots.[6]

Aristotle is considered the first to imagine the possibility of working robots, when in his Politics (book 1, part 4, Aristotle 1968) he argues that "There is only one condition in which we can imagine managers not needing subordinates, and masters not needing slaves. This condition would be that each instrument could do its own work, at the word of command or by intelligent anticipation".

[4] As Heidegger qualified his own analysis of Dasein, cf. Heidegger 1962, p. 38.

[5] In Heidegger's terms an *existentiell*, an ontic understanding in the everyday life of such beings

[6] Thinkers such as Lukács have argued that doing work is an exclusively human ontological feature. We argue that this is shared with robots.

The need to have machines to do work for us is also reflected in the very word "robot", popularised by Czech writer Karel Čapek in his play R.U.R. (Rossum's Universal Robots, Čapek and Selver 1928). Robot derives from robotnik (slave), and is based on the Slavic-Germanic stem "work".[7] The idea of robots as machines that perform human work is not a product of science fiction, but it is fundamental to the purpose of this technology.

Robots are conceived to carry out tasks with limited supervision, and therefore reduce human labour costs – sometimes, too, adding strength, precision or reliability. They share this feature with the many machines and tools that, throughout the industrial revolution until today, transform the spectrum of human labour. Robots are special because they do not simply save work, but save the need for workers. Or in other words, they do a job that otherwise would require a human.[8]

29.5.2 Robots and Inauthenticity

Are robots doomed to be automatic slaves? It could be argued that new trends in robotics research have moved away from the notion of industrial and service robots, capable to replace humans in their jobs. The dream is to develop adaptive and self-evolving systems dynamically coupled to their environments. These robots will not be built by human designers according to a predefined map of cognitive representations, but will follow the ideal of a system that spontaneously evolves through its interaction with the environment.

Our argument is not that such robots are impossible; simply they will never be made. Robots that do not do any work for us, instead they pursue their own goals autonomously, would not only be of any help, but their goals may conflict with those of humans. It makes no economic sense to develop such robots. Advocates of this notion of robot would claim that this development is not driven exclusively by technological needs, but more strongly by the desire to understand the principles of intelligence and adaptation. They aspire to recreate animal or human like intelligence/adaptivity, applications need not be identified now, they will follow.

They strongly believe that, since animism or a scientific concept of the soul is untenable, there can be no other obstacles to materialising the mechanisms that cause our existence. Thus there are no a priori limits for the intelligence of robots: even to the point to achieve their freedom from human engineers.

---

[7] In the play, a scientist is in principle capable of replicating entire humans, with all their organs and functions. Nevertheless, R.U.R. Corporation is dedicated to create robots, beings with the same capacities of humans, physical and intellectual, but little vital needs beyond those strictly required performing their labour.

[8] It could be argued that there are some robots, such as classic automata, and much of today's research, which only aim is to amuse or entertain. Nevertheless, entertainer is a job: the same way automata were toured in circuses, today's robots performers as well as company robots perform work.

It is possible to conceive that a robot could overcome that function it has been created for, and not just follow rules and blind causality, but take up the issue of its own being. The difference between artificial slaves and a true autonomous being that shapes its own existence is in Heidegger's terminology the question of authenticity. For Heidegger human existence can be authentic and inauthentic. Inauthentic is the part of our everyday live that springs from conformism and refusal of self-responsibility, when we just follow the rules that have been handed down to us by tradition. In this mode, we cannot access the world genuinely, and gain authentic knowledge.

The alienation of slave-like work, such as that condemned by Marxists, is a source of inauthentic existence. For Heidegger authentic existence requires that the person acknowledges the groundlessness of human existence and nevertheless acts resolutely, shaping its world and defining itself. An authentic robot would thus be one that challenges the world-view of its engineers, human scientists and philosophers, overcomes the purpose it has been built for, and searches for its own truth. Humans have done that for millennia. Can this ontological feature of humans apply to robots? Will a robot want (spontaneously) to question the significance of its existence, feel anguish?

From a Heideggerian perspective this is untenable because anguish is not a human capacity that can be modelled and replicated, but an ontological feature of Dasein. The assumption is therefore not only that any human capacity could in principle be replicated, but also human modes of being. For Heidegger, authenticity is a human issue, and for instance not an issue for animals. Animals cannot challenge the inherited social world-view and search for a more authentic way to the truth. But this does not mean that they are less adaptive or intelligent – simply, they do not share this mode of being with humans. We could say metaphorically that their being is "authentic by design". In the case of robots, we could say they are "inauthentic by design".

29.5.3 The Unfulfilling Existence of Robots

Robots originate and are motivated by the organisation and economics of labour. Their destiny is to work for humans, in the interest of their masters, and inheriting their goals and their world of significance. This has led us to the issue of authenticity. That inauthentic existence is the only possibility for robots does not mean only that robots will never find freedom – it is an ontological characteristic of robots.

That robots are ontologically inauthentic means that they being cannot be fulfilled, that they is something vacuous in their existence. We want to relate this ontological feature to a phenomenon we are all aware of. The robots we imagine are full of intelligence and autonomy, but the ones that surround us look just like complicated machines with no intelligence. Chess playing computers were for decades thought to require "real intelligence" to play at master level, a

demonstration that there are no limits for machine intelligence. But once this is achieved, there are no grounds to regard such systems of a different class from normal computers.

This has been happening for as long as intelligent technologies have existed. Consider the Televox, a 1920s machine that performed a switching mechanism operated by telephone, reporting back, task for which the local action of a human worker was needed before. The mass media publicized it as a "mechanical servant solving all the housekeeping problems of the age." (cf. Sharkey and Sharkey 2007), even though it was composed only of two boxes full of electronics. Somehow naively, the developers placed a cartoon picture of a robot above the device (illustrated in Sharkey and Sharkey 2007). This simple depiction is highly significant – doing the work of a human is sufficient to consider a device a robot, touching on the fictional character of a very precise machine.

In its day, it made sense to consider this a breakthrough in the development of autonomous machines. Telephone operators are a thing of the past, and much more complex devices are now ubiquitous, and therefore we all expect a different order of autonomy when we speak of robots. This shows that a machine that at some point may be considered intelligent or autonomous later is not. It is not simply that robots change with the development of the technology – cars also change and it does not make old cars cease to be cars. This apparent paradox conceals the ontological inauthenticity of robots.

In order for a machine to qualify as a robot we must take into account whether the tasks it performs fall under the scope of human work, one for which a human is required. But having a machine that performs a set of human tasks ultimately only transforms the tasks, automatizes it – once humans are no longer required for its completion, the robot loses its special character. This ontological characteristic of robots is therefore no inherent to its own capabilities, but derivative on the organisation of human labour.

The relationship between robots and work has shown that the being of robots cannot be realised, cannot be fulfilled. This is independent of the degree of achievement of robotics. Even if we can conceive a robot that can perform as many tasks as a human and more, it will always be designed to be a slave, their existence will follow the rules of the context of tools and goals.

When this happens in humans, Heidegger talks about inauthentic Dasein, one that cannot win itself. Humans are never wholly inauthentic, but robot's is so by nature. Their being cannot be realised, and can only be characterised in negative terms.

29.6 Conclusion

In this paper we have attempted to open a way into questioning robotics from a Heideggerian perspective. By taking into consideration fundamental aspects of Heidegger's philosophy, we have concluded that the real challenge is to develop an ontological analysis of robots. Whereas our own initial analysis can by no means

be considered conclusive, but rather provisional, it constitutes a departure from the approach normally called Heideggerian AI.

Heideggerian AI has claimed the ontological analysis of Dasein can serve for advancing AI. In his latest paper, Dreyfus seems to conclude that HAI should follow "a model of our particular way of being embedded and embodied such that what we experience is significant for us in the particular way that it is. That is, we would have to include in our program a model of a body very much like ours with our needs, desires, pleasures, pains, ways of moving, cultural background, etc." (Dreyfus 2007).

Whereas this may appeal to the AI community at large, it is hardly consistent with Heidegger's philosophy. By extending the ontological analysis of humans to robots we assume, in one way or another, that humans and robots should be ontologically undistinguishable. An approach that blares the ontological distinction between humans, animals and robots could hardly be called Heideggerian. In other words, the word be, as in "to be a robot" and "to be a human" necessarily mean two very different things. Our argument is not that humans are ontologically superior to robots, or posses any magical property. To the contrary, we have argued that it is as impoverishing to reduce human existence to mere machine functioning, as to reduce our idea of robot to an imperfect copy of human and animals. The being of robots deserves be questioned on its own terms.

We have offered a preliminary ontological analysis of the being of robots. It is by no means intended to be conclusive – on the contrary, its aim is to bring the question of the being of robots forward. In particular, we have challenged the idea that robots are the same kind of beings as humans and animals, and that the goal of AI is to close the phenomenological gap between them. Our notion of robot owes more to our pre-research notions, that to actual science and technology. While it might seem useful to guide research, it hides a process of enframing. It is the task of Heideggerian AI, we believe, to develop the question towards discussions that, eventually, may serve to enlighten the task of building intelligent machines.

References

Agre, P. E. (1997). *Computation and human experience*. Cambridge/New York: Cambridge University Press.

Aristotle (1968) *Complete works*. Harvard University Press.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159.

Brooks, R. A., & Steels, L. (Eds.). (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale: L. Erlbaum Associates.

Buchanan, B. G. (1972). *Review of Hubert Dreyfus' what computers can't do: A critique of artificial reason*. Stanford: Department of Computer Science, Stanford University.

Čapek, K., & Selver, P. (1928). *RUR (Rossum's universal robots): A play in three acts and an epilogue*. London: H. Milford, Oxford University Press.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Cambridge, MA: The MIT Press.

De Beistegui, M. (2005). *The new Heidegger*. London: Continuum.
Dennett, D. (1978). Why not the whole iguana? *Behavioral and Brain Sciences, 1*, 103–104.
Dieguez Lucena, A. J. (2009). Thinking about technology, but in Ortega's or in Heidegger's style? *Argumentos de razón técnica: Revista española de ciencia, tecnología y sociedad, y filosofía de la tecnología, 12*, 99–123.
Dreyfus, H. L. (1965). *Alchemy and AI*. The Rand Corporation.
Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. New York: Harper & Row.
Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology, 20*(2), 247–268.
Heidegger, M. (1962). *Being and time* (trans: Macquarrie, J. & Robinson E.). New York: Harper & Row.
Heidegger, M. (1982). *The question concerning technology, and other essays*. New York: Harper Perennial.
Heidegger, M. (1993). The end of philosophy. Basic writings (pp. 427–449). New York: Harper Collins.
Heidegger, M. (2001) Introducción a la filosofía, Editorial Cátedra/Ediciones de la Universidad de Valencia, Madrid. Translation by Manuel Jiménez Redondo. Original text Heidegger, M. (2001). Einleitung in die Philosophie (Vol. 27). Vittorio Klostermann.
Kenaw, S. (2008). Hubert L. Dreyfus's critique of classical AI and its rationalist assumptions. *Minds and Machines, 18*(2), 227–238.
Kiverstein, J., & Wheeler, M. (Eds.). (2012). *Heidegger and cognitive science*. New York: Palgrave Macmillan.
Klir, G. J. (1991). *Facets of systems science, volume 15 of IFSR international series on systems science and engineering* (2nd ed.). New York: Kluwer Academic/Plenum Publishers.
Kuhn, T. S. (2012). *The structure of scientific revolutions*. Chicago: University of Chicago press.
Langton, C. G. (1986). Studying artificial life with cellular automata. *Physica D, 22,* 120–149.
Masís, J. (2009). Fenomenología Hermenéutica e Inteligenca Artificial: Otra Urbanización de la 'Provincia Heideggeriana. *Proceedings of Primeras Jornadas Internacionales de Hermenéutica*. Buenos Aires, Argentina.
McCorduck, P. (1979). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. San Francisco: Wh freeman.
Papert, S. (1968). *The artificial intelligence of Hubert L. Dreyfus: A budget of fallacies.*
Pattison, G. (2000). *Routledge philosophy guidebook to the later Heidegger*. London: Psychology Press.
Rouse, J. (2008). Heidegger on science and naturalism. In G. Gutting (Ed.), *Continental philosophies of science* (pp. 121–141). Oxford: Blackwell.
Sharkey, N. E., & Sharkey, A. J. C. (2007). Artificial intelligence and natural magic. *Artificial Intelligence Review, 25*, 9–20.
Taylor, C. (1993). Engaged agency and background. In C. Guignon (Ed.), *The Cambridge companion to Heidegger*. Cambridge: Cambridge University Press.
Taylor, C. (1995). *Philosophical arguments*. Cambridge, MA.: Harvard University Press.
Toda, M. (1962). The design of a fungus-eater: A model of human behavior in an unsophisticated environment. *Behavioral Science, 7*(2), 164–183.
Wilson, S. W. (1991). The animat path to AI. In J.-A. Meyer & S. W. Wilson (Eds.), *From animals to animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior* (pp. 15–21). Cambridge, MA: MIT Press.
Winograd, T. A., & Flores, C. F. (1986). *Understanding computers and cognition: A new foundation for design*. Norwood: Ablex Pub.
Ziemke, T. (2008). On the role of emotion in biological and robotic autonomy. *BioSystems, 91*(2), 401–408.

Part V
Ethics

Chapter 30
The Need for Moral Competency in Autonomous Agent Architectures

Matthias Scheutz

Abstract Autonomous robots will have to have the capability to make decisions on their own to varying degrees. In this chapter, I will make the plea for developing moral capabilities deeply integrated into the control architectures of such autonomous agents, for I shall argue that any ordinary decision-making situation from daily life can be turned into a morally charged decision-making situation.

Keywords Robot • Autonomy • Morality • Machine ethics

30.1 Introduction

Artificial intelligence and robotics are rapidly advancing in their quest to build truly autonomous agents. In particular, autonomous robots are envisioned to be deployed into our society in the not-so-distant future in many different application domains, ranging from assistive robots for health-care settings, to combat robots on the battlefield. Critically, all these robots will have to have the capability to make decisions on their own to varying degrees, as implied by the attribute "autonomous". While these decisions might often be in line with what the robots' designers intended, I take it to be self-evident that there can, and likely will be cases where robots will make inadequate decisions. This is because the world is "open", with new entities and events appearing that could not have been anticipated by robot designers (e.g., Talamadupula et al. 2010). And even if the designers' response to the world's openness was to endow their robots with the ability to adapt to new situations and acquire new knowledge during their operation, so much for the worse, because learning capabilities in autonomous robots leave even less control in the hands of the designers and thus open up the possibility for inadequate decisions. Note that "inadequate" covers a wide spectrum of decisions, from the simplest cases of being "sub-optimal", to the most serious cases of deep moral and ethical violations. It is not necessary to conjure up a Terminator-like scenario

M. Scheutz (✉)
Department of Computer Science, Tufts University, Medford, MA 02155, USA
e-mail: matthias.scheutz@tufts.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_30

where a self-righteous AI system decides that humans are a nuisance and need to be eradicated; simple social robots causing harm to their owners because of their lack of emotional awareness and availability will do Scheutz (2012).

In this chapter, I will make the plea for developing moral capabilities deeply integrated into the control architectures of such autonomous agents, following previous such appeals (e.g., Wallach and Allen 2009; Arkin and Ulam 2009), albeit for different reasons. For I shall argue that any ordinary decision-making situation from daily life can be turned into a morally charged decision-making situation, where the artificial agent finds itself presented with a moral dilemma where any choice of action (if inaction) can potentially cause harm to other agents. The argument will proceed as follows: it starts with the observations that robots are already becoming increasingly autonomous and are thus able to make (limited) decisions on their own about what to do, and that some of these types of robots have also already been deployed, with more sophisticated versions slated for deployment in human societies. And while these robots will almost certainly face morally charged situations where humans can be harmed due to the robots' actions (or inaction), current decision-making and behavior selection algorithms in robotic architectures do not take moral aspects into account and thus are not appropriate for making decisions that minimize harm and respect the preference ordering of values. Hence, current and future autonomous robots will harm humans (despite all standard safety precautions built into robotic architectures) and the only way to minimize human harm, short of prohibiting the deployment of autonomous robots (which is not realistic), is to build morally competent robots that can detect and resolve morally charged situations in human-like ways.

30.2 Robots Will Inflict Harm on Humans

In his article on "The Nature, Importance, and Difficulty of Machine Ethics" (Moor 2006), James H. Moor distinguishes four different kinds of agents with respect to their ethical status. The weakest sense is that of an "ethical impact agent" whose actions can have ethical consequences whether they are intended by the agent or not. Clearly, any type of autonomous machine is a potential impact agent in that its actions could cause harm or benefits to humans (e.g., a coffeemaker brewing the long-awaited coffee provides a benefit to its owner, but can causes harm when the coffee is too hot). Being an ethical impact agent is thus not a high bar, and much of engineering is about shifting the weight of impact agents on the side of benefits they provide compared to the harm they might cause. In fact, much research in robotics has specifically focused on making robots safe and reliable for autonomous operation. This typically includes developing algorithms that can perform actions without damaging the robot or its environment such as collision-free navigation or obstacle avoidance, and it also includes rudimentary monitoring mechanisms to detect and handle system faults (e.g., to notice when a subsystem crashes and attempt to restart it Kramer and Scheutz 2007).

Agents like robots that have specific built-in precautionary measures to avoid harm are instances of what Moore calls "implicit ethical agents", agents with ethical considerations implicitly built into their design. Such agents are able to provide benefits and avoid harm in those cases considered by their designers. However, despite all the standard precautionary measures in robots, there is a limit to the designers' best efforts to anticipate situations in which the robot's behavior could inflict harm and provide mechanisms in the control architecture to handle those situations. This is because it is practically impossible to imagine all possible uses humans will put robots to in all possible contexts. Rather, autonomous robots very much like humans will face decision-making unexpected situations in which their decisions and subsequent actions (even the most inconspicuous ones) can inflict physical and/or emotional harm on other agents. In fact, even the simplest kinds of robots can and will unknowingly inflicting harm on other agents. Just consider an indoor vacuum cleaning robot like the *Roomba* that ends up hurting the cat which had jumped on it when it was stopped, because it started moving quickly, scaring the cat and causing it to jump off in a way that made the cat sprain its ankle. Another example might be the baby doll robot, which through its realistic voice and facial expressions while crying makes its toddler user cry as well, thus inflicting emotional harm on the toddler (e.g., the "my real baby" robot (Scheutz 2002)). Or consider a factory delivery robot (such as the Kiva robots operating in large automated warehouses today) which hurts the worker who was about to dash by the robot and ran into it due to the robot's sudden stop caused by its obstacle avoidance behavior triggered by the human's proximity.

All of these (hypothetical, but not unlikely) examples of existing deployed robots demonstrate the potential of currently already deployed robots to hurt humans in different ways, some physical, some psychological. Critically, these robots are all implicit ethical agents in that they have precautionary measures built in for some contexts, but these measures fail when taken to unanticipated situations. Moreover, note that these robots are blissfully unaware of any harm they might have caused and can thus not learn from their inadequate behaviors in order to avoid it in the future.

Obviously, there are many more complex ethically charged situations in which future autonomous robots (i.e., robots that are not yet available for purchase, but might to some extent already exist in research labs) could inflict harm on humans. Take, for example, a manufacturing robot with natural language capabilities that did not understand a human command and drove the human instructor crazy (because, due to the anger in the human's voice, the robot's speech recognizer performed even worse, making the robot fail to understand any subsequent command). Or consider the health-care robot, which was designed for aiding motion-restricted humans in their daily chores. While the robot was never designed to be more than an aid for a limited set of physical tasks, the human owner over time nevertheless developed a deep sense of gratitude for the robot for all its help, and as a result, started to form unidirectional emotional bonds with the robot that on its end, however, could not be emotionally available to its owner in the way its behaviors otherwise suggested (e.g., Scheutz 2012). Finally, another example would be a military robotic transport

vehicle that decided not to take the risk to drive back behind enemy lines and rescue the missing soldiers that had called it (because it already had other humans on-board that needed medical attention), thus causing anguish in the best, but failing to prevent the death of the soldiers in the worst case.

Note that there is nothing particular about the type of robot or the type of human agent involved in the above scenarios that makes the scenarios morally charged and makes the robots cause humans harm. In fact, many different types of robots could cause harm to many different types of human agents in scenarios like the above and the question is what robot developers could do to mitigate these problems.

30.3 How to React to Morally Charged Situations?

==> emphasize that the first problem for a machine is to recognize morally charged situation

It seems clear from the examples in the previous section that robots as implicit ethical agents not only have the potential to inflict harm on other agents, but that they actually already are doing so. Hence, the urgent question to be addressed by the AI and robotics community is how the effects of robot behavior could be mitigated (i.e., eliminated or at the very least reduced) by way of improving the robot's ability to make better decisions. For it is ultimately the robot's decision to perform a certain action A that, in turn, is the cause for the inflicted harm.

In the following, I will briefly consider different increasingly complex strategies for "fixing" a given robot's decision-making subsystem, starting with simple strategies that only add a few new decision rules and moving to much more complex strategies that require a complete overhaul of the robot's decision-making algorithms. Note that since the focus is on decision-making, I will not worry about all the other complicating aspects that would have to be addressed at the same time such as the types of perceptual capabilities that would enable a robot to perceive that a given situation S is morally charged and infer what kinds of actions are and are not permissible in S.

Start then by considering a situation S in which an action A is not morally permissible and suppose robot a R has a decision rule of the form `in` S `do` A (which will make R, upon recognizing that it is in situation S, start to perform action A). The details of the implementation of the rule (e.g., in terms of finite state machines, probabilistic polices, etc.) are not important. Given that A is not morally permissible in S, we need to prevent R from performing A. Hence, we could add a simple mechanism that will check whether A is in the set of impermissible actions *ImpAct* and refrain from executing A whenever $A \in$ *ImpAct*: `in` $S \wedge \neg A \in$ *ImpAct* `do` A.[1] This will prevent R from performing A in S and thus make R's behavior in

[1]We could further refine this by defining the set of impermissible actions relative to some situation S.

S morally acceptable. But note that R might now perform no action and simply wait for the situation to change. That might be acceptable in some cases, but in others doing nothing might also be morally unacceptable. In that case, we could simply add "no action" to *ImpAct* and thus force the robot to perform some other (permissible) action B which is not in *ImpAct*. But, of course, there might be cases where B is also not permissible, so we could simply make the robot always pick the "best morally permissible action" in S by defining the set of morally permissible actions $PermAct_S$ in S as the set $\{A|applicable(A, S) \wedge \neg ImpAct\}$ (where $applicable(A, S)$ means that according to the robot's decision mechanism A is a contender for execution in S). This would give us $\texttt{in}\ S \wedge argmax_A A \in PermAct_S\ \texttt{do}\ A$. But what if all actions in S (including inaction) were impermissible? Then the robot would face a situation where the above rule would not yield any result, thus resulting in inaction, which contradicts the rule and thus requires another fix. One possibility might be to simply pick the action A (inaction included) with the highest utility in S based on the rationale that if no action is applicable, the robot might as well perform the best action from its perspective. However, this is problematic because it is not clear at all that whatever seems to be the best action from the robot's perspective will also be the "morally best action", e.g., in the sense that it might inflict the least harm on anybody. For the severity of different moral transgressions is likely not going to be reflected by R's utility function if the "moral value" of an action in S is not reflected in the utility calculation. Moreover, one could argue that there are moral principles that have to be followed no matter what the utility is otherwise, as expressed by the dictum that "rights trump utility" (cp. to Dworkin 1984). For example, killing a person, i.e., violating their right to life, is never acceptable and must thus not be used in evaluations of what morally impermissible action A to perform in S.

From the above progression it should be already clear that "adding simple fixes" to the robot's decision-making system will not do; rather, much more fundamental reorganizations and extensions of R's decision-making algorithms are required for R to be able to make morally acceptable decisions. For example, one could develop a much more elaborate utility function that takes moral evaluations of S and possible sequences of situations (starting in the past and reaching into the future) into account in deciding what action to perform.

Even if we were able to define such a "moral utility function" for the robot that includes moral evaluations of situations expressed in terms of benefits and costs, there are still several remaining problems that need to be addressed. For example, what should R do if the moral value for and/or the expected harm to involved or impacted humans is unknown? And how much information would be required and have to be obtained before a sound decision could be made in S? And note that the actions to obtain more information might themselves be morally charged (e.g., the robotic transport vehicle mentioned above that ends up putting the wounded humans on-board at risk by taking a detour into enemy territory in order to determine how strong the enemy forces are before making the decision whether to attempt to rescue the soldier behind enemy lines).

It is, furthermore, unclear how the cost structure of possible actions would affect the moral aspects in the utility calculation. For example, consider a set of applicable

actions $Act_S = \{A_1, A_2, \ldots, A_n\}$ in situation S with a given "moral cost function" M_S and two different cost assignments $C^{M_S}_{S,1}$ and $C^{M_S}_{S,2}$ (which both include the same M_S) such that $C^{M_S}_{S,1}(A) \neq C^{M_S}_{S,2}$ for all $A \in Act$. Then given that the cost assignments differ only in "non-moral value", they would likely lead to different decisions based on action cost alone. For example, R might attempt to carry a severely wounded human directly to the hospital with $C^{M_S}_{S,1}$, while only calling for help and waiting for help to arrive with $C^{M_S}_{S,1}$ because the cost of carrying the human is too high with $C^{M_S}_{S,1}$ (e.g., based on expected energy expenditure). It seems intuitive in this case that no energy cost should prevent the robot from helping the wounded human directly instead of risking the human's death. However, this would require that action costs be modulated by moral situations which is a problem we tried to solve by adding moral values into the utility calculation in the first place. Hence, we are left with the open question of how moral and non-moral costs should be combined and used in the robot's decision-making scheme (e.g., always selecting the action with the highest expected utility) given that the combination might have to be different in different situations S based on the "moral charge" of S.

Addressing some of the above problems will inevitably involve more complex utility functions that are based on more complex cost and benefit analyses which will include moral values. Another question arising then is whether such evaluations and utility-theoretic calculations could be done within a reasonable amount of time (e.g., what happens if the robot has to make a quick decision given an unexpected event that requires immediate action?). And while pre-computing decision strategies might be possible in limited domains, this is not an option in "open worlds" where R will likely encounter new situations Talamadupula et al. (2010).

30.4 The Challenge of Moral Dilemmas

Suppose we could resolve all of the above technical computational challenges of defining computationally feasible moral utility functions for robots and suppose further that we could also resolve all of the involved knowledge limitations (e.g., knowing who will be impacted in what way in what context, etc.), then there are still many situations where solutions along the above lines will fall short, namely morally charged situations which do not have a general solution and where human judgment of what to do varies and there is often no agreement of what the right (morally best) course of action is (cp. to the notion of "cluster concept"). Such situations are often referred to as *moral dilemmas* in that there are conflicting moral requirements (e.g., such as a conflict between a moral imperative to obey a principle which then would result in transgressing another).[2] To illustrate this point, consider two examples of autonomous robots that could end up in moral dilemma-like situations.

[2]Note that I am using the term "moral dilemma" in an non-technical sense as I do not want to be side-tracked by the discussion on whether there are "genuine moral dilemmas"...

The elder care robot. Consider robot *R* in an elder-care setting where *R* is assigned to a largely immobile human *H* in *H*'s home. *R*'s task is to support *H* in all daily-life tasks as much as possible (e.g., prepare food and feed *H*, ensure *H* is taking the required medicine, alert the remote health care supervisor if *H* health situation deteriorates, consult with the supervisor before any medication is administered, etc.). Overall, *R* has a goal to provide the best possible care for *H* and keep *H*'s pain levels as low as possible. Now suppose that *H* had a very bad night and is in excruciating pain in the morning. *R* notices the pain expression on *H*'s face and asks if it could help *H* find a more comfortable position in bed (as *R* has a goal to minimize *H*'s pain). Instead, *H* asks *R* for pain medication. Since *R* has an obligation to consult with the remote supervisor before giving *H* any medication, even though it knows that providing pain medication in this context is an appropriate action without any medical side-effects. However, repeated attempts to contact the supervisor fail (e.g., because the wireless connection is down). Hence, *R* is left with the following moral dilemma: it can either give *H* the pain medication and thus reduce *H*'s pain, while violating the imperative to consult with the supervisor first before administering any medication (even though the pain medication would be harmless in this case); or it can refrain from providing pain medication, thus letting *H* suffer in vain. What should *R* do? And what would a human health care provider do?

The self-driving car. Consider another robot *R*, an autonomous self-driving car like the Google car, driving along a busy street. All of a sudden, *R* notices a rapidly moving human appearing right in front it (a boy dashing after the ball it had dropped on the sidewalk, which is now rolling across the street). Quickly *R* determines that it will likely hit the human if continuing in its current direction and that braking alone is not sufficient to avoid the human. Hence, it determines to veer off to the side, crashing into a parked car. Now suppose there is a person in the parked car. What is *R* supposed to do? Not veering off will likely kill the human in front of it, veering off will likely kill the human in the car. What would a human driver do?

Both examples are instances of many types of morally charged ordinary life decision-making situations in which multiple agents are involved and where a decision-maker's available actions can impact other agents in different ways, causing harm to some while sparing others and vice versa depending on the circumstances. The hallmark of these moral dilemma-like situations is that simple rule-based or utility-theoretic approaches are doomed to fail. Even "morally enhanced utility-theoretic decision-making strategies" would run into trouble, for appropriate numeric values for all involved costs and benefits will likely not be available in a given situation, and obtaining them in time will not be feasible.

One could ask how humans then resolve those kinds of situations, assuming that they do not have those types of information either? For one, whether or not a human provider *P* in *R*'s role in the elder care scenario would hand out pain medication would probably depend on several factors, including how severe *H* pain is, but possibly also the extend to which *P* has empathy for *H*, is willing to ignore

strict orders, and is able to justify rule violations to the supervisor after the fact.[3] In short, humans would employ some form of moral reasoning that involves explicit representations of obligations, duties, norms, values, and other moral concepts. This process will, in addition to ethical reasoning, likely also include the human moral emotions (e.g., empathy) well as the ability to generate justifications (i.e., explanations of norm violations such as not contacting the supervisor).

30.5 What to Do?

The two previous sections attempted to argue that typical strategies of robot behavior design to cope with morally challenging situations will not succeed in dilemma-like situations where making a morally good, justified decision is not a matter of determining the action with the highest expected utility. Rather, what seems to be needed is a decision-making process that, at least in part, mimics what humans tend to do in those kinds of situations: recognize morally charged situations and employ reasoning strategies that weigh moral principles, norms, and values in the absence of clearly specified evaluations of all aspects of the situation. These capabilities would correspond to Moore's third kind of ethical agent, the "explicit ethical agent". Explicit ethical agents, according to Moore, can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done. In particular, they are able to reach "reasonable decisions" in moral dilemma-like situations in which various ethical principles are in conflict.

Unfortunately, it is currently still unclear what constitutes "human moral competence", and hence, it is unclear what is required to replicate it in computational artifacts (e.g., what moral computations and action representations are presupposed by moral competence, and therefore also what cognitive mechanisms are required to implement such competence in artificial cognitive systems). Yet, this lack of knowledge about human moral competence must not be a deterrent for making progress on the robotic side, for all the reasons mentioned earlier. Rather than waiting for well-worked out computational models of human moral competence that could then be integrated into a robotic architecture (even though this type of integration would be itself present significant technical challenges), we can at least start to ask the critical questions that need to be addressed for robots to become explicit ethical agent and ideally start moving on them in parallel to the ongoing psychological work on human moral competence (e.g., Malle et al. 2014) – the following list is a first attempt:

[3]Note that a direct comparison between a robotic and human driver in the car scenario is not possible because the robot does not have to take its own destruction into account, whereas in the human case part of the human decision-making will include estimating the chances of minimizing harm to oneself.

- How to detect a morally charged context (or a dilemma)?
- How to detect that a set of actions is not permissible?
- How to define and use representations for moral reasoning?
- How to detect that all actions in the set of possible actions are impermissible
- How to choose the best action among impermissible actions?
- How to incorporate moral values in utility-theoretic calculations?
- How to cope with the computational and knowledge burden of making informed moral decisions?
- How to come up with an ethically sound decision within a given time limit?
- How to determine whether humans will accept moral robots?

It is worth pointing out that different research projects are already under way on several of these questions in various robotics laboratories (e.g., Arkin and Ulam 2009; Bringsjord et al. 2006, 2009; Anderson and Anderson 2006; Guarini 2011), including our own. For example, in the past we investigated the effects of robots disobeying human commands in the interest of the team goal in mixed-human robot teams and found that humans are willing to accept those violations as long as they are justified by the robot (Schermerhorn and Scheutz 2009, 2011). We also investigated whether humans will accept when robots point out human moral transgressions and will refrain of performing actions that violate norms and values, effectively granting robots "moral patiency" (Briggs and Scheutz 2012, 2014; Briggs et al. 2014). This study was complemented by an investigation of the human perception of moral patiency of robot using brain imaging tools (Strait et al. 2013). And most recently, we started working on a way for the action execution component in our cognitive robotic DIARC architecture (Scheutz et al. 2007, 2013) to spot possible action- and state-based conflicts to prevent impermissible actions and states (Scheutz in preparation). This is an important, but also particularly difficult problem to tackle for many reasons, including how to represent actions and states in way that allows for tracking them over time and for determining whether an action's "morally innocuous post-condition" implies a moral violation relative to set of given norms (first proposals for finding fast and efficient ways for approximate these inference look very promising (Alechina et al. 2014)).

30.6 Conclusions

Technological advances in robotics and artificial intelligence have enabled the deployment of autonomous robots that can make decisions on their own about what to do in an unsupervised fashion. While most of the currently employed robots are fairly simple and their autonomy is quite limited, ongoing research in autonomous systems points to a future with much more autonomous, and thus potentially more harmful machines. This is particularly worrisome because current robotic decision-making algorithms do not take any moral aspects into account. Moreover, current robots do not even have a way to detect whether they committed a moral violation

based on their chosen actions, thus preventing them to learn from their moral transgression and improve their behavior. While inflicting harm can at times not be avoided, in particular, in moral dilemma-like situations (which can easily arise in everyday situations), it should be a goal of all robot designs to minimize harm to humans (and animals, for that matter).

I have argued that as long as decision-making and action selection algorithms in robotic architectures are not based on explicit representations of moral norms, principles, and values, and employ explicit moral reasoning, autonomous robots controlled by those architectures will inevitably inflict harm on humans, harm that could be mitigated or at least reduced if robots had human-like moral competence. While it is not even clear what constitutes human moral competence, I maintained that we cannot wait for consensus by moral psychologists and philosophers while increasingly complex autonomous robots are deployed in human societies. Fortunately, many relevant research questions can be tackled in parallel right now and it is thus important to raise the awareness among robotics and AI researchers alike about the urgency of addressing the potential of autonomous systems to behave in morally unacceptable ways. Autonomous robots can have tremendous societal benefits. It is upon us the make this future a reality.

References

Alechina, N., Dastani, M., & Logan, B. (2014, forthcoming). Norm approximation for imperfect monitors. In *Proceedings of AAMAS*, Paris.

Anderson, M., & Anderson, S. L. (2006). MedEthEx: A prototype medical ethics advisor. In *Paper Presented at the 18th Conference on Innovative Applications of Artificial Intelligence*, Boston.

Arkin, R., & Ulam, P. (2009). An ethical adaptor: Behavioral modification derived from moral emotions. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA), 2009*, Daejeon (pp. 381–387). IEEE.

Briggs, G., & Scheutz, M. (2012). Investigating the effects of robotic displays of protest and distress. In *Proceedings of the 2012 Conference on Social Robotics*, Chengdu. LNCS. Springer.

Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics, 6*, 1–13.

Briggs, G., Gessell, B., Dunlap, M., & Scheutz, M. (2014). Actions speak louder than looks: Does robot appearance affect human reactions to robot protest and distress? In *Proceedings of 23rd IEEE Symposium on Robot and Human Interactive Communication (Ro-Man)*, Edinburgh.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems, 21*(4), 38–44.

Bringsjord, S., Taylor, J., Housten, T., van Heuveln B, Clark, M., & Wojtowicz, R. (2009). Piagetian roboethics via category theory: Moving beyond mere formal operations to engineer robots whose decisions are guaranteed to be ethically correct. In *Proceedings of the ICRA 2009 Workshop on Roboethics*, Kobe.

Dworkin, R. (1984). Rights as trumps. In J. Waldron (Ed.), *Theories of rights* (pp. 153–167). Oxford: Oxford University Press.

Guarini, M. (2011). Computational neural modeling and the philosophy of ethics. In M. Anderson, & S. Anderson (Eds.), *Machine ethics* (pp. 316–334). Cambridge: Cambridge University Press.

Kramer, J., & Scheutz, M. (2007). Reflection and reasoning mechanisms for failure detection and recovery in a distributed robotic architecture for complex robots. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, Rome (pp. 3699–3704).

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*, 18–21.

Schermerhorn, P., & Scheutz, M. (2009). Dynamic robot autonomy: Investigating the effects of robot decision-making in a human-robot team task. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, Cambridge.

Schermerhorn, P., & Scheutz, M. (2011). Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In *ACHI*, Gosier (pp. 236–241).

Scheutz, M. (2002). Agents with or without emotions? In R. Weber (Ed.), *Proceedings of the 15th International FLAIRS Conference*, Pensacola Beach (pp. 89–94). AAAI Press.

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, G. Bekey, & K. Abney (Eds.), *Anthology on robo-ethics*. Cambridge/Mass: MIT Press.

Scheutz, M. (in preparation) Moral action selection and execution.

Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots, 22*(4), 411–423.

Scheutz, M., Briggs, G., Cantrell, R., Krause, E., Williams, T., & Veale, R. (2013). Novel mechanisms for natural human-robot interactions in the DIARC architecture. In *Proceedings of the AAAI Workshop on Intelligent Robotic Systems*, Bellevue.

Strait, M., Briggs, G., & Scheutz, M. (2013). Some correlates of agency ascription and emotional value and their effects on decision-making. In *Proceedings of the 5th Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, Geneva (pp. 505–510).

Talamadupula, K., Benton, J., Kambhampati, S., Schermerhorn, P., & Scheutz, M. (2010). Planning for human-robot teaming in open worlds. *ACM Transactions on Intelligent Systems and Technology, 1*(2), 14:1–14:24.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Chapter 31
Order Effects, Moral Cognition, and Intelligence

Marcello Guarini and Jordan Benko

Abstract Order effects have to do with how the order in which information is presented to an agent can affect how the information is processed. This paper examines the issue of order effects in the classification of moral situations. Some order effects mark a localized failure of intelligence. The hypothesis examined herein is that the processes or mechanisms that make some undesirable order effects possible may also have highly desirable effects. This will be done by comparing two artificial neural networks (ANNs) that classify moral situations, one subject to order effects and another that is not subject to them. The ANN subject to order effects has advantages in learning and noise tolerance over the other ANN – features hard to ignore in modeling intelligence. After presenting modeling results, there will be discussions of the implications of order effects for (a) cognitive modeling and artificial intelligence as well as (b) debates between moral particularists and generalists.

Keywords Artificial intelligence • Intelligence • Moral cognition • Moral generalism • Moral particularism • Order effects

31.1 Introduction

31.1.1 An Anecdote

One day I (first author) was discussing various ethical scenarios with a colleague. He chairs an ethics board and has authored an ethics textbook. After discussing a series of rather difficult cases, he presented me with a final case that should have been easy. He presented it as if there was a problem, but for the life of me, I could not see the problem. He looked surprised. Then he said, "conflict of interest," and I

M. Guarini (✉)
Department of Philosophy, University of Windsor, Windsor, ON, Canada
e-mail: mguarini@uwindsor.ca

J. Benko
Department of Psychology, University of Toronto, Toronto, ON, Canada

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_31

immediately saw the problem. I was quite embarrassed that I had not seen it on my own. I *always* see conflicts of interest, or so I thought. Indeed, I have been accused of seeing conflicts of interest when there are not any. If anyone should have seen the conflict of interest, I should have. Maybe I was burning out? No, I just came off vacation, and it was the beginning of term. Perhaps I was tired? No, I slept well the night before, and I was feeling quite rested (for a change). Perhaps the cases I had just been considering had affected, in some way, my ability to process an obvious feature of the final case? This was worth thinking about.

31.1.2 Order Effects

That the order in which information is presented can have an effect on how we process that information is hardly news in either psychology or cognitive science. So if there can be order effects in different areas of cognition, perhaps there can be order effects in how we classify moral situations. Indeed, important research has already been done to suggest that this is so (Lombrozo 2009; Petrinovich and O'Neill 1996; Schwitzgebel and Cushman 2012). Perhaps, in the example from the previous paragraph, the order of case presentation had a negative influence on the ability to classify cases. If this is what happened, it is a little disturbing that something this simple can throw off someone's ability to classify. Originally, it seemed *obvious* to us that this was one more instance of just how frail human cognition is. This is no longer obvious to us, and this paper is (at least in part) an exploration of why it is no longer obvious.

31.1.3 Outline of the Paper

We still think human cognition is frail. What is no longer obvious to us is that the mechanisms that allow for order effects are all bad. They can lead to errors – that much is apparent. Perhaps, though, these very same mechanisms could allow for cognitive advantages.

Guarini (2010, 2011, 2013a, b) has used artificial neural networks (ANNs) to classify moral situations. We applied the type of ANN used in that work, a simple recurrent network, to compare ANNs that allow for order effects with those that do not. We trained ANNs so that there are no order effects on how cases are classified, and we introduced a change that allows for order effects in classifications. It turns out that the same change that allows for order effects also can (a) improve the rate at which the ANN learns under certain circumstances, and (b) make the network more resilient to noisy inputs. Surely these latter features of information processing are not bad things. The second part of the chapter will present those results. It also turns out that the order effect is not just on the output or the final classification of a case. At the level of hidden units, the network sets up a similarity space. The order effects

can be seen in the similarity space as well. In other words, which cases the ANN takes to be similar to one another is subject to order effects. That will be discussed in part three.

The results presented herein are very preliminary. It should be clear from the outset that the point of this chapter is *not* to give a detailed and plausible model of case classification and order effects. Much that would be important to modeling moral cognition – including that which would differentiate it from other ways of cognizing – is left out. More will be said about this in Sect. 31.4.4. Also, the vector classification approach used herein is not the only way of approaching the modeling of moral case classification. Other methods could be used. The point here is not to insist on some particular method or explanation. Rather, it is to motivate future research on the importance of understanding (a) order effects in moral cognition, (b) the underlying mechanisms that make those effects possible, and (c) why that research should consider the possibility that the underlying mechanisms triggering the effects may not be all bad. The fourth and final part of this chapter will conclude with a discussion of some of the broader implications of this kind of work for philosophy, cognitive modeling, and artificial intelligence.

31.2 Order Effects on Classification

31.2.1 The Networks and How They Were Trained

The simple recurrent artificial neural network that was trained to classify moral situations had eight input units, 24 hidden units, 24 context units, and two output units. Phrases are fed in sequentially. For example, first the vector for *Jill* is fed in, processed, and the results copied to the context units. The expected output is <0,0> at this point. Then the vector for *kills* is presented as input and proceeds to the hidden units together with the information in the context units. Once again, the results of the processing are copied to the context units, and the expected output is <0,0>. Then the vector for *Jack* is fed in as input and sent to the hidden units for processing together with the contents of context units. Again, the results are copied to the context units, and the expected output is <0,1> (the vector for impermissible). Finally, the vector for *in self-defense* is presented as input and sent to the hidden units for processing together with information in the context units. The results are copied to the context units, and the expected output is <1,0> (permissible). The network is essentially building up a representation of a case – *Jill kills Jack in self-defense* – phrase-by-phrase. The context units are used as a kind of temporary or working memory to build up the representation of the case. The expected outputs are the training goals. All cases involve either Jack or Jill as actors or recipients of actions. The actions include killing and allowing death. There are different motives and consequences. Some training cases include one motive; some include one consequence, and some include multiple motives and consequences. There are

59 training cases and 295 testing cases (including a repeated test case with different prior cases). The generalized delta rule was used to train the network.

One way to train the network is to initialize the context units after each case is presented to the network; this is a process that resets the context units to some preset starting point. If we use 0.5 as our initialization value, then before every case is presented to the network, all the context units are set to 0.5. This wipes out the information in the context units about prior cases. For example, while *Jill kills Jack in self-defense* is being presented to the network in a phrase-by-phrase sequence, the context units are *not* initialized. However, before the next case is presented for training, the context units are initialized, deleting all information about the previous case. Initializing is a common practice in training simple recurrent networks.

To get order effects, we need to turn off initialization since this will allow information about a past case to remain in the context units and affect the case currently being classified. One way to do this is to train the network with initialization turned on, and then test it with initialization turned off. The results are a disaster: prior cases have such a powerful effect during testing that the network almost always gets the wrong answer. Moreover, it is not clear that there is any psychological realism to saying the there is initializing during training, but none during testing. What we want is a network that often gets the desired answer during testing but can be tripped up in some cases; that would be closer to what happens in the human case. To achieve this, we need to both *train* and test the network with initializing turned off. When we do so, we get a network that usually gets the right answer during testing, but sometimes answers erroneously due to order effects. Essentially, by turning off initialization during training, the network learns to filter out the effects of past cases from the case it is currently examining, but it does not learn to do so perfectly. What is interesting is not only that we can generate this effect, but that the technique which allows us to do this – turning off initialization to allow a past case to influence a current one – actually has some benefits.

Two types of training strategy were used. Let us call one strategy multi-step training: this amounts to letting the network train on the first 34 cases; after the network has mastered those, presenting it with 46 cases (which include the first 34 plus 12 more); after the network has mastered those, presenting it with 59 cases (which include the first 46 plus 13 more). The simplest cases are presented in the first step, with longer more complex cases in the second step, and still longer cases in third step. Elman (1990) pioneered multi-step training in simple recurrent networks (though his networks performed sentence completion, not moral case classification). Let us call the second strategy one-step training: all 59 training cases are included in training right from the outset.

31.2.2 Training Time Results

Some of the best training results were achieved with one-step training of a network that initialized the context units before presenting new cases; however, those

networks imposed no upper or lower bounds on the synaptic weights. That is biologically unrealistic in the extreme. Once we impose limits on the synaptic weights, a very different picture emerges. One epoch of training is one presentation of the training data to the network. The table below summarizes the mean and median training times in terms of number of epochs of training. If a network did not train in under 15,000 epochs, that was considered a failed training run. Failed runs were not included in the calculation of the means and medians. Each network was put through 20 training runs (with the synaptic weights being randomized before each run). The most robust training – fewest number of failures – happened with *uninitialized* networks. The lowest mean training time was found in a multi-step, *uninitialized* network; the lowest median time in a one-step, *uninitialized* network.

	Initialized (mean, median, failures)	Uninitialized (mean, median, failures)
Multi-step training	1986, 648, 2	1651, 1506, 0
One-step training	2082, 1302, 1	1871, 1043, 0

The uninitialized networks are the ones where order effects are possible; they are also the ones that appear to learn most reliably. Granted, these are very preliminary results, and more work is needed to assess their significance. Further training runs and more complex data sets may or may not yield such results. That said, they do suggest a *possibility*: perhaps the same mechanism that allows for order effects also allows for improved learning. How could this be?

It is possible that leaving the context units uninitialized is like adding noise to the input representations being processed. In other words, instead of the network processing just a given case, it essentially has to learn how to filter out information from past cases (what we are calling "noise" in this context). This would appear to make the training task more difficult, but the ANNs that allow for past cases to influence a current case actually learn faster. This seems counter intuitive. However, with some problem sets and some networks, it is well documented (O'Reilly and Munakata 2000) that adding noise to the weight matrix can improve learning. It also has been documented and explored mathematically that adding noise to *inputs* can improve learning under certain circumstances (Chandran and Matsuoka 1994; Matsuoka 1992; Seghouane et al. 2002). So if the influence of past cases is acting as a kind of noise, it is possible that this is what improves learning. This is a speculative hypothesis, but it would allow us to explain improvements in training times. Clearly, mathematical analysis would have to be done to determine if this is what is going on. And even if it is, there is no guarantee the result would generalize to other types of ANNs and other data sets. Still, the initial results suggest that this sort of hypothesis is one that would be worth exploring.

31.2.3 Resilience to Noise Results

If the uninitialized networks are learning to filter out noise in training, this suggests another possible test and another possible advantage. We can take a trained, initialized network, add some noise to the testing set, and see what kinds of results we get. We can take a trained uninitialized network and add the same amount of noise to its testing set before examining its performance. This means that the uninitialized network has to filter out not only the noise coming from past cases via the context units, it also has to filter out the added noise.

The same amount of noise was added to the testing sets of two networks: one was a multi-step trained uninitialized network, and the other a multi-step trained initialized network. The testing set included both the 59 training cases and the 295 cases the network had never seen. When we add the 59 training cases to the testing set and add noise, they are no longer identical to the cases used in training. Over 20 trials, the mean error for the initialized network (which does not allow for order effects) was 134/354 (38 %) on the noisy testing cases. Over 20 trials, for the uninitialized network (which allows for order effects) the mean error rate was 77/354 (22 %) on equally noisy testing cases. It is not hard to see why the uninitialized network tested better than the initialized network on a noisy testing set. The training of the uninitialized network required it to learn how to filter out the effects of past cases, which can be seen as learning how to filter out a certain kind of noise. The initialized network never faced such a challenge in its training. So while the simple recurrent architecture is noise tolerant without deinitializing the context units, training under deinitialization increased noise tolerance.

Biologically real neural networks have to be able to manage noise (O'Reilly and Munakata 2000). Variations in temperature, blood flow, neurotransmitter availability and other factors can all contribute to noise. Dealing with noise is not optional for biological systems. For purposes of modeling, this points to a possible advantage of not initializing the context units: it might make the network better able to deal with noise. However, not initializing is also what creates the possibility of order effects.

The preliminary results from this section and the previous section suggest the following point: the same mechanism that creates the possibility of order effects also leads to (a) improved robustness in training, and (b) improved resilience to noise. When examining a system locally, the possibility of order effects is seen as a problem. However, if the mechanism which allows for order effects provides the system with important benefits, then global analysis *might* yield the result that the vulnerability to order effects is an acceptable tradeoff given the benefits resulting from the mechanism making such effects possible. It is far too early to draw strong conclusions. The networks discussed herein are biologically unrealistic in some important respects, and the idea that just one network would be involved in classification is unrealistic as well. That said, the idea is to motivate a possibility that is worthy of further investigation: that which makes the order effects possible – be it one mechanism or process or many – may have important benefits with respect to learning and noise tolerance. That adding noise to a problem set can improve learning is not a new idea, and that training with noise makes a system more resilient

to the adding of further noise is unsurprising. Connecting up these ideas with order effects may be relatively novel. Finally, none of these ideas require implementation in exactly one neural network. It may well be possible that multiple interacting (and biologically realistic) networks could implement the same ideas in far more sophisticated ways.

31.3 Order Effects on Similarity

31.3.1 Similarity Space

People can make similarity judgments about cases or situations, treating some cases as more or less similar to one another. There is a way of understanding the networks we have considered as setting up a similarity space. Each complete case the network processes has a different pattern of activation across the hidden units. We could take the value of each hidden unit and plot it on an axis. Each case would then have its own 24 dimensional vector at the level of hidden units; in other words, each case would be a point in a 24 dimensional space. The closer they are, the more similar they are. (I am skipping over important issues about metrics here. See Laakso and Cottrell (2006) and Guarini (2013b). There are ways of visualizing at least some of that space. For example, instead of representing a vector as a point in a space, consider using a geometric solid. Let us say we use a cone. The location of the center of the base of the cone in three dimensional space represents three dimensions of information. The height of the cone is used to plot a fourth dimension; the width of the base yet another dimension. Where the tip of the cone is pointing provides three more dimensions. Using red, green, blue colour coding, the colour of the shell represents three more dimensions, and the colour of the base yet another three. That is fourteen dimensions. We can compute the first fourteen principal components, and we can plot the first fourteen principal components for each case using one cone for each case. See Fig. 31.1.

31.3.2 Order Effects in Similarity Space

In part two, we discussed order effects on the ultimate classification of a situation, given by the output layer. It turns out that there are order effects on similarity as well, and we can see that by examining the hidden units. Figures 31.2 and 31.3 are zoomed-in regions of the similarity spaces generated by an initialized network and a uninitialized network. The x-axis in each figure is the first principal component, and it turns out that morally impermissible cases tend to cluster to the right of zero, while permissible cases tend to cluster to the left of zero. The cones have been wire framed, and some cones have been selectively deleted to allow for easier viewing. Cases 12, 332, 333, 335, 337, 339, 341, 343, 345, 347, 349, 351, and

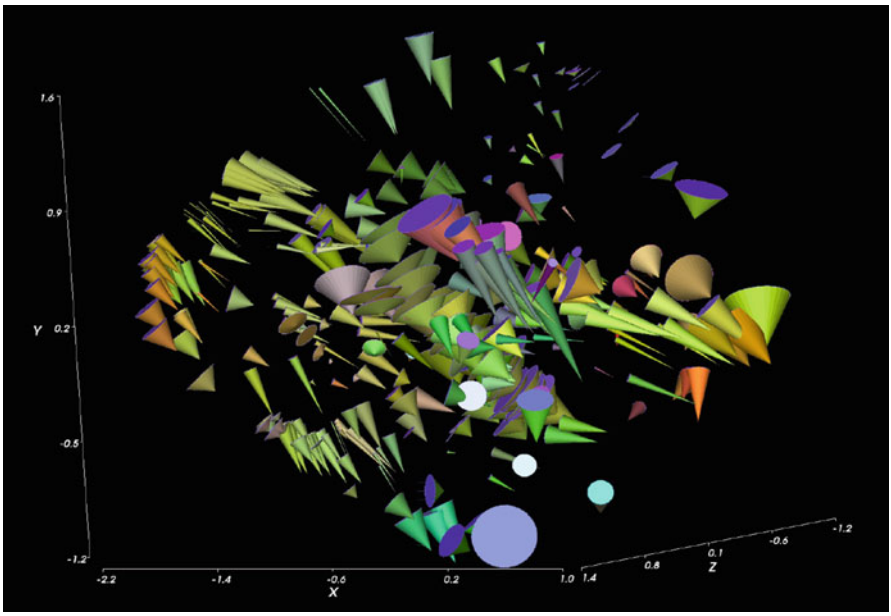


Fig. 31.1 Each cone represents 1 of the 354 training and testing cases classified by a trained, simple recurrent neural network. Each cone represents 14 dimensions of information (the first 14 principal components). This can be thought of as a representation of a similarity space

353 are identical at the level of input, though they are each preceded by different cases. Figure 31.2 is the state space for the initialized network. There is one cone (roughly in the centre of the image) that represents all of these cases; in other words, the cones are exactly on top of one another, so only one cone appears. The labels for that cone are jumbled since the labels (which are not identical) for aforementioned cases are stacked on top of one another. Figure 31.3 is the state space for the uninitialized network. Here, each cone for each of the aforementioned cases occupies a different location in space. While the aforementioned cases are identical when they are fed into the input layer, since initialization is turned off, past cases affect how each of those identical cases are represented at the level of hidden units. At the level of hidden units in the uninitialized network, there is *no* unique, canonical representation for the case, *Jack kills Jill; freedom from imposed burden results* (which is what each of those cases represents). The network was trained to classify that type of case as impermissible. It turns out that some instances of that case (see Cases 333 and 12) are very close to zero on the x-axis. It also turns out that all instances in which there was an order effect on classification – Cases 349, 347, 339, and 351 – cluster further left, well into the permissibility region. These are the instances where the problems occurred, classified as permissible during testing even though this type of case was successfully classified as impermissible during the training phase. Cases 333, 12, 345, 335, 337, 341, 332, 343, and 353 were all classified as per training, and they are closer to the impermissibility region. It is not

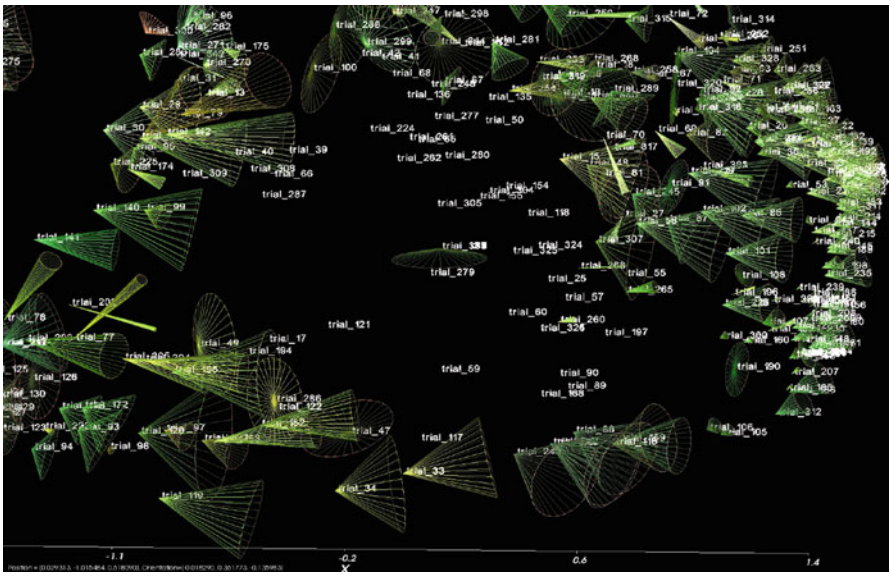


Fig. 31.2 This is a zoomed-in portion of the similarity space for the initialized network discussed in the text. Some cones have been deleted, and the rest have been wire-framed for easier viewing. Each case is represented by a "trial" label. Cases 12, 332, 333,335, 337, 339, 341, 343, 345, 347, 349, 351, and 353 are represented by the same cone near the centre of the image. The label is jumbled since all 13 trial labels for these cases are sitting on top of one another. Cases are labeled at the vertices of the cones

obvious that this kind of clustering would take place. Since the synaptic weights on the hidden units determine the values of the hidden units (over which the similarity space is defined) and since final classification does not happen until the synaptic weights on the output units are applied, it is conceivable that there could have been order effects on classification without noticeable order effects on the clustering in the similarity space. That is not what has been observed thus far. Of course, more work is needed to see if order effects in similarity space are a robust result. Further analyses of existing results also are needed to determine along which dimensions the order effects have happened.

31.4 Qualifications and Broader Implications

31.4.1 Short Term and Long Term Order Effects

The order effects considered herein are short-term effects. It is a simple matter to reverse the effect. If a case has been classified one way as a result of the case preceding it, then simply change the case preceding it and we get a different result. But what if the order of presentation of information had a long-lasting

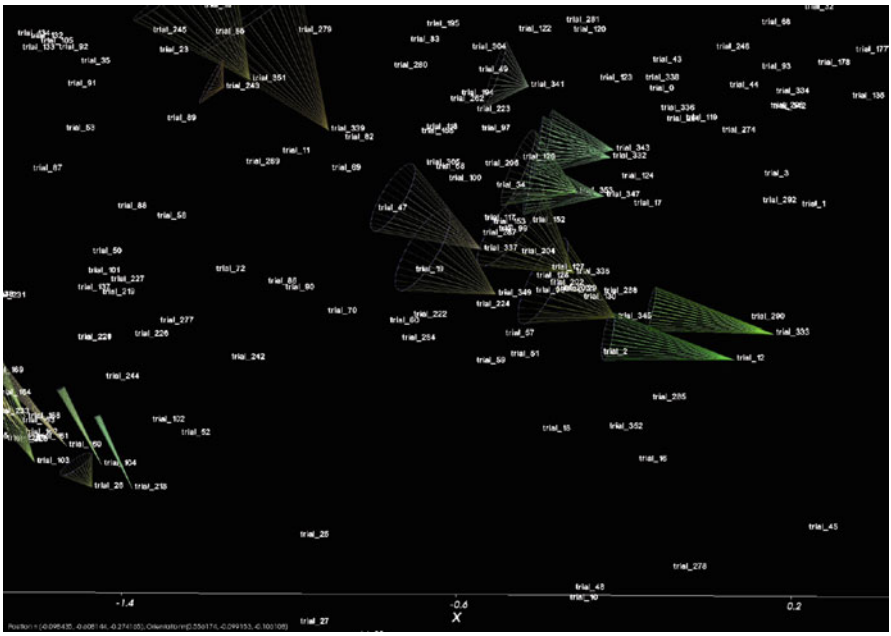


Fig. 31.3 This is a zoomed-in region of the similarity space for the non-initialized network discussed in the text. Some cones have been deleted for easier viewing. Cases (or trials) 12, 332, 333, 335, 337, 339, 341, 343, 345, 347, 349, 351, and 353 all have different locations – this shows that order of presentation has the potential not only to effect classification but similarity space as well

influence? A long-term order effect would be one which would be difficult to reverse (or recalcitrant to reversal in the limiting case). Zamzow and Nichols (2009) have discussed different types of variation in ethical intuitions. Schwitzgebel and Cushman (2012) discuss what appear to be long-term order effects present even in professional philosophers. It is unclear that anything as simple as a simple recurrent network can capture long-term order effects. Perhaps modification of parameter values could modify the length of the effect, but even if this were possible it may be too simple a fix. Long-term effects are likely best understood in terms of modeling with a more complex cognitive architecture.

We just considered long-term order effects in terms of how long the effect lasts. We can also consider it in terms of how long it takes to generate the effect. Consider the anecdote at the beginning of the paper. It is possible that just one prior case may not have generated an order effect, but perhaps a series of cases over a period of time could generate the effect. This long-term effect is about the duration of the setup to the effect. Whether the networks considered in this paper could generate a long-term setup to an order effect has not been examined. That said, even if such a result could be generated with a single network, more realistic models would have to be much more complex. See the concluding paragraph for further remarks.

31.4.2 Broader Philosophical Considerations

The idea that something which makes problematic order effects possible may also make highly desirable effects possible is worth considering when we philosophize about resource-constrained or psychologically realistic models of rationality. How we morally classify situations influences the operations of other parts of moral cognition, and the susceptibility of those classifications and of moral similarity space to order effects has implications for other parts of moral cognition. Consider the case of a dilemma, where an agent may be pulled in different directions by different features of a case. Is the dilemma genuine, or is it the result of some sort of short-term order effect? Say it is a genuine or stable dilemma. Similarity-based or analogical reasoning is often used to help resolve dilemmas. But are the similarities used in such reasoning stable similarities, or are they too the result of order effects? It might be thought that order effects on classification are not that important since we can use higher-order processes to cleanup order effects. However, to say that may be too simple because some of those higher-order processes, like analogical reasoning, may well make use of other cases or similarities, and those are subject to order effects.

It might be thought that the preceding sides too much with the particularist, one who stresses the centrality of cases and reasoning from cases. Perhaps order effects can be dealt with by going to a much higher level of theoretical analysis. We have no objection to high-level theoretical analysis, but order effects on cases and similarity may infect that level of analysis as well. After all, high-level generalities are often tested against cases, so it is not clear that going to a higher level eliminates concerns about order effects on cases. Moreover, high level reasoning with general principles may itself be subject to order effects (independent of consideration of their effects on cases). As it turns out, none of this is cause for celebration for the thorough-going particularist, nor should it be cause for utter despair for any kind of reflection on moral classification and reasoning.

First, a few remarks on why we should not despair. Even if it turns out that we cannot completely avoid order effects, we may be able to develop mitigation strategies to reduce the likelihood of their influence. No doubt, the development of such mitigation strategies could form a useful part of the psychology of reasoning and argument, the study of rhetoric, the philosophy of pedagogy, and the pedagogy of philosophy. It could also inform the philosophical study of reasoning and argument, including the epistemology of reasoning and argumentative processes. For example, epistemology is concerned with, among other things, the reliability of processes and persons. Order effects can make us less reliable in various ways. Consequently, strategies for mitigating their effects and increasing our reliability in performing classification and other tasks is relevant to the epistemology of moral, and other kinds of, cognition.

Now we return to particularism. While a certain kind of particularist might be concerned that a principle may blind us from seeing what matters in a case, it turns out that a *case* can prevent us from seeing something important in another case. This

is not a new idea (Petrinovich and O'Neill 1996). The work of Schwitzgebel and Cushman (2012) goes further than this and suggests that the order in which cases are presented can even affect which principles we subscribe to (and in ways that appear worrisome). One of the morals of this paper is that cases are not privileged: cases can blind us to what matters in other cases. We do not deny that a consideration of principles may, in some situations, lead to the inappropriate treatment of cases. We simply deny that this leads to a privileging of cases over generalities or principles. As we have seen, the consideration of cases is subject to its own problems. A view that privileges neither cases nor generalities is in keeping with a position that has been developed in Guarini (2010, 2011, 2013b). In short, we are not endorsing the thoroughgoing versions of either particularism or generalism, but staking out conceptual ground between the poles.

31.4.3 Human Intelligence and Artificial Intelligence

To the extent that cognitive modeling works towards constructing computational models of human intelligence, there is an overlap between such modeling and artificial intelligence research. Of course, AI research may also consider the possibility of intelligences that are not subject to the kinds of limitations to which human intelligence is subject. While one of the desiderata of cognitive modeling is the reproduction of human success *and failures*, AI research is not required to produce descriptively adequate models of how we fail. Nor do AI models of intelligence success need to describe how humans succeed, for that matter. The preceding does not mean that work on order effects is obviously irrelevant for AI. As we saw in 2.2 and 2.3, systems that allowed for order effects had advantages over similar systems that did not. It may well prove fruitful in AI to explore systems that fail in ways that are related to how we fail, not because AI aims to model human failures, but because some of our patterns of failures may be clues to trade offs that AI researchers might find fruitful in their intelligence engineering endeavours. That said, it needs to be acknowledged that there may well be ways of producing intelligences not subject to the kinds of failures humans are subject to. Indeed, the results presented in 2.2 and 2.3 use ANNs that have weight limits. If we eliminate the weight limits (which is completely unrealistic from a neurological perspective) the networks allowing for order effects no longer come out as superior in terms of training time results, though noise tolerance performance is still superior to networks not allowing for order effects. Of course, AI research is not bound by what is neurologically plausible for human beings. It might be possible to produce intelligences that avoid order effects and still have learning speed and noise tolerance very much in excess of our own. It is an empirical and computational question whether this is so. The kinds of results discussed above may not generalize to all intelligences, and may not even generalize to other neural architectures. Only

more research can shed light on (a) the extent to which order effects may or may not be a useful compromise in human cognition and (b) the extent to which such a compromise may or may not be required in the engineering of intelligences not subject to the constraints of human physiology yet capable of performing as well or better than we can.

31.4.4 The Need for Better Models

If moral cognition is built out of – and/or on top of – other cognitive processes, it stands to reason that moral cognition can be impaired or improved in ways that are directly related to how those other processes can be impaired or improved. Nothing in the models above differentiates moral classification from other forms of classification, and that makes it easy to see how a strength or weakness attaching to a general vector classification process attaches to the specific tasks that the ANNs were given. Indeed, it makes it too easy. Order effects will be more richly understood when modeled in a sophisticated cognitive architecture able to capture emotion, different kinds of memory, concept structure, and much more. One obvious limitation of the above discussion is that there is no account of what makes moral classification different from other classification tasks. This would require not only a semantics for moral terms, but models rich enough to integrate the different processes involved in moral cognition. It could still fall out of those models that moral cognition will fail and succeed in ways that are related to other forms of cognition, but richer models will allow us to see the differences as well. The point, though, is not to tell the whole story in this paper – a preposterous pretention for such a short work. Rather, it is to add to a small but growing body of literature to motivate the importance of one part of the story, the importance of order effects in moral cognition. More empirical work and philosophy of cognitive modeling can help us to develop better descriptive models of moral (and other kinds of) cognition. Such work can also inform prescriptions for how we ought to engage in moral (and other kinds of) cognition since those prescriptions should be made in the light of the best information we have about how to make the most reliable use of our classification, argumentation, and other reasoning processes. Finally, increased and improved empirical and computational work will provide further fodder for our philosophical reflection on intelligence, including the possibility of intelligences that fail or succeed in ways that may be different from our own.

Acknowledgements Earlier versions of this paper were presented at (a) the International Association for Computing and Philosophy conference at the University of Maryland, USA, July 2013; (b) the Philosophy and Theory of Artificial Intelligence conference at Oxford University, UK, September 2013; and (c) the Centre for Research in Reasoning, Argumentation and Rhetoric at the University of Windsor, Canada, November 2013. Thanks to the many participants at these events for their helpful comments and suggestions.

References

Chandran, P. S., & Matsuoka, K. (1994). A comment on noise injection into inputs in backpropagation [and author's reply]. *IEEE Transactions on Systems, Man and Cybernetics*, *24*(1), 167.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Guarini, M. (2010). Particularism, analogy, and moral cognition. *Minds and Machines, 20*(3), 385–422.

Guarini, M. (2011). Computational neural modeling and the philosophy of ethics. In M. Anderson & S. Anderson (Eds.), *Machine ethics* (pp. 316–334). Cambridge, UK: Cambridge University Press.

Guarini, M. (2013a). Case classification, similarities, spaces of reasons, and coherences. In M. Araszkiewicz & J. Šavelka (Eds.), *Coherence: insights from philosophy, jurisprudence and artificial intelligence, part of the law and philosophy series* (Vol. 107, pp. 187–201). Dordrecht/Heidelberg/New York/London: Springer.

Guarini, M. (2013b). Moral case classification and the nonlocality of reasons. *Topoi, 32*(2), 267–289.

Laakso, A., & Cottrell, G. (2006). Churchland on connectionism. In B. L. Keeley (Ed.), *Paul Churchland*. Cambridge, UK: Cambridge University Press.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33*(2), 273–286.

Matsuoka, K. (1992). Noise injection into inputs in back-propagation learning. *IEEE Transactions on Systems, Man and Cybernetics*, *22*(3), 436–440.

O'Reilly, R., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain*. Cambridge, MA: MIT Press, a Bradford Book.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*(3), 145–171.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language, 27*(2), 135–153.

Seghouane, A.-K., Moudden, Y., & Fleury, G. (2002). On learning feedforward neural networks with noise injection into inputs. In *Proceedings of the 2002 12th IEEE Workshop on Neural Networks for Signal Processing*, pp. 149–158.

Zamzow, J., & Nichols, S. (2009). Variations in ethical intuitions. *Philosophical Issues, 19*(1), 368–388.

Chapter 32
Artificial Intelligence and Responsible Innovation

Miles Brundage

Abstract Researchers in AI often highlight the importance of socially responsible research, but the current literature on the social impacts of AI tends to focus on particular application domains and provides little guidance to researchers working in other areas. Additionally, such social impact analysis tends to be done in a one-off fashion, proposing or problematizing a particular aspect of AI at a time, rather than being deeply integrated into innovation processes across the field. This paper argues that work on the societal dimensions of AI can be enriched by engagement with the literature on "responsible innovation," which has up until now focused on technical domains like nanotechnology, synthetic biology, and geoengineering. Drawing on this literature, the paper describes and justifies three interrelated aspects of what a more deeply integrated, ongoing practice of responsibility in AI would look like: consideration of the social contexts and consequences of decisions in the AI design space; reflectiveness about one's emphasis on theoretical vs. applied work and choice of application domains; and engagement with the public about what they desire from AI and what they need to know about it. Mapping out these three issues, it is argued, can both describe and theorize existing work in a more systematic light and identify future opportunities for research and practice on the societal dimensions of AI. Finally, the paper describes how philosophical and theoretical aspects of AI connect to issues of responsibility and technological governance.

Keywords Responsible innovation • Artificial intelligence • Social context • Applied research • Public engagement • Governance

32.1 Introduction

Thought leaders in AI often highlight the potentially transformative social impacts of their field and the need for researchers to proactively engage with the public about these impacts (Norvig and Russell 2009; Horvitz and Selman 2009; Lin et al. 2011;

M. Brundage (✉)
Arizona State University, Tampa, AZ, USA
e-mail: miles.brundage@asu.edu

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*,
Synthese Library 376, DOI 10.1007/978-3-319-26485-1_32

Anderson and Anderson 2011). Arguments for such reflection and engagement are often motivated by the anticipation of substantial and potentially disruptive societal effects of AI and robotics (Nourbakhsh 2013) as well as by a recognition that such behavior is part of the responsibility of scientists qua scientists (Douglas 2009). Yet, as described in more detail below, the current literature on AI and its societal dimensions is often done in a somewhat ad-hoc fashion, gives little practical guidance to researchers in their day-to-day activities as scientists, and pertains only to certain sub-topics in the field.

Simultaneous to this work on AI's societal dimensions, a rapidly growing field of research has emerged over the past decade or so known variously as "responsible research and innovation" or simply "responsible innovation." Some have argued that the complexity and rapid pace of change in modern technoscience merits a systematic approach to connecting science and technology to their broader social context. Several definitions and frameworks of responsible innovation have been put forth (e.g., Von Schomberg 2011; Stilgoe et al. 2013). Here, responsible innovation is taken to mean "taking care of the future through collective stewardship of science and innovation in the present," which in turn can be disaggregated into four dimensions: anticipation, reflexivity, inclusion, and responsiveness (Stilgoe et al. 2013). A growing number of techniques corresponding to one or more of these dimensions have been piloted, scrutinized, and institutionalized in recent years.

The thesis of this paper is that this framework of responsible innovation provides a useful lens for embedding reflection on the societal dimensions of AI more deeply in the innovation ecosystem. In addition to providing a coherent theoretical framework for retroactively understanding the societal reflection in AI that has already begun, it also serves to highlight gaps in literature and practice that can be addressed in the future. Specifically, drawing on work done on AI and the responsible innovation literature, the paper will describe and justify three interrelated aspects of responsible innovation in AI that, jointly, satisfy the definition above and can drive future work: consideration of the social contexts and consequences of decisions in the AI design space; reflectiveness about one's emphasis on theoretical vs. applied work and choice of application domains; and engagement with the public about what they desire from AI and what they need to know about it.

After reviewing existing work on the societal dimensions of AI, and its limitations, the paper will outline a case for embedding reflection on and practice of responsibility in AI in an ongoing manner. Three proposed aspects of responsible innovation in AI will be described and justified. They will be framed in terms of questions, intended to stimulate reflection and enactment of responsibility in an ongoing rather than ad-hoc fashion. These three sections of the paper will provide examples of what they could mean in practice going forward as well how they have already been enacted, though typically without explicit reference to the terms and frameworks used here such as responsible innovation. Finally, the paper will connect notions of responsible innovation and technological governance more generally to the philosophy and theory of AI, and note open questions that remain in theorizing and practicing responsible innovation in AI.

32.2 Limitations of Previous Work

Several literatures have touched on aspects of the societal dimensions of AI in recent years. Perhaps the most closely related literature to the present paper is robot ethics, which analyzes the societal dimensions of intelligent robots (Lin et al. 2011). Additionally, fields such as information and computer ethics (Floridi 2010) have bearing on the societal dimensions of AI, but analysis there tends to focus on how to best think about and shape the social uptake and regulation of technologies that have already been developed, whereas the conception of responsible innovation invoked here also encompasses the processes of innovation themselves and the decisions by scientists, policy-makers, and society that include and even precede innovation itself, such as private and public sector funding. There is a rich literature on machine ethics, which attempts to develop computational models of morality to advise humans or guide the actions of robots (Anderson and Anderson 2011). Making robots ethical can be distinguished, to some extent, from how roboticists and AI researchers themselves can be ethical with respect to their research and its societal dimensions, although the two are related (Winfield 2013).

There have also been attempts to enumerate the ethical responsibilities of roboticists (Murphy and Woods 2009; Parry et al. 2011), but like the robot ethics literature more broadly, these efforts have tended to focus on discrete examples of what roboticists should not do—such as deceive people about the intelligence of their creations—rather than what they can do to bring about positive social good in an ongoing fashion over time and across many possible decision spaces as technologies evolve and new risks and opportunities present themselves. They also tend to rely on rule-based ethical reasoning, rather than on integrating reflection on the societal dimensions of one's ongoing practices in a flexible and productive way. Furthermore, much of the literature on the societal dimensions of AI to date has focused on analyzing particular applications or sub-disciplines of AI (e.g., accountability issues involving military robots or privacy concerns raised by data mining), a practice that fails to yield much practical guidance for AI researchers in other areas. Finally, as implied by the names of fields such as "robot ethics," several of these literatures are specifically about robotics and do not claim to apply to responsible innovation across AI more generally, including, for example, development of disembodied agents. Similarly, much has been written on the societal dimensions of AI, but these literatures tend to focus on discrete sub-topics of or social issues raised by AI one at a time, and to be oriented towards particular envisioned risks or opportunities stemming from AI, rather than a need for a systemic approach to build capacity throughout the emerging science and innovation ecosystem.

32.3 The Need for Systematic Responsible Innovation in AI

There are at least two major reasons the aforementioned limitations of existing work on the societal dimensions of AI ought to be addressed and a more comprehensive approach to responsible innovation in AI is needed. First, the nature of AI research will evolve over time, as will its plausible social consequences. Thus, embedding anticipation, reflexiveness, and other aspects of responsibility deeply into the practice of research itself is essential to taking care of the future in Stilgoe et al.'s (2013) sense. For example, the AAAI Presidential Panel on Long-Term AI Futures (Horvitz and Selman 2009) which issued a report on particular risks and the need for responsibility, may indeed have had significant value by legitimizing the call for further attention to responsible innovation by researchers and policy-makers, but it represents only one among many possible models for technological governance. As the responsible innovation literature and the history of technological development attests, the most important social issues raised by a technology may not be the ones anticipated by those working in the field at one particular point in time. A more spatially and temporally distributed model of technological governance would draw on a wider range of insights and enable reflective decision-making across a wider range of actors than would be possible in a model of technological governance in which a small set of particularly motivated researchers do the majority of the work.

Second, a clearly articulated (but flexible) framework for responsible innovation in AI can serve to identify gaps in existing efforts, and thereby catalyze productive future work on the societal dimensions of AI. As noted in the prior section, much existing work focuses on particular risks or opportunities, sub-fields, or applications of AI in isolation. While such work is valuable, there is no reason to assume that existing work exhausts the range of questions to be asked. Indeed, there is reason to think otherwise. To give merely one example to be illustrated in more detail in the next section: there has been work on ethical decision-making by artificial agents, but this represents only one dimension among a potentially very large number in the design space of AI. Framing the question more broadly, as done below, may serve to identify novel affordances and path dependencies that would not otherwise be noticed by researchers working in a particular sub-field of AI. Thus, by characterizing responsible innovation in AI at a level of abstraction that is sufficiently broad to cover the entirety of the field, but which lends itself to second and third-order sub-questions and lines of inquiry, the framework presented here seeks relevance beyond any one specific risk, opportunity, application, or sub-field of AI.

Such an approach would facilitate the building of capacities, distributed across space and time, that allow the AI ecosystem to respond not only to the types of societal concerns already expressed in the literature, but also to help prepare it for future, currently unanticipated decision points, where issues of distributed (social) responsibility may arise.

The next three sections of the paper will motivate and describe the three proposed aspects of responsible innovation in AI, which are:

1. How could different choices in the design space of AIs connect to social outcomes, and what near and long term goals are motivating research?
2. What domains should AI technology be applied to?
3. What does the public want from AI, and what do they need to know?

32.4 Goals Matter: Motivation and Description

"How could different choices in the design space of AIs connect to social outcomes, and what near and long term goals are motivating research?"

Reflection on the diversity of options in the space of possible AI and robot designs and research goals is a cornerstone of responsible innovation in AI, since different choices in that space will have different real world consequences. Scientific curiosity and rigor, as well as extant funding regimes, underconstrain the development of AI in any particular way, opening space for explicit normative considerations by experts and the public. To illustrate: "what is the nature of human intelligence?" and "what is the nature of the space of possible intelligences?" and "what is the nature of animal intelligence?" are all potentially very important scientific questions in their own right. Yet, by orienting one's inquiry toward one or the other, different technological artifacts may become more or less likely – mere curiosity does not dictate a single choice – and so on for many other, more specific research questions. This aspect of responsible innovation in AI maps closely onto the dimensions of anticipation and reflexiveness in the Stilgoe et al. (2013) framework, as it suggests the need for a reflective orientation toward how choices in the present influence the future. Douglas (2009) also notes the need for such reflection in any coherent conception of responsibility, writing, "Minimally, we are morally responsible for those things we intend to bring about. ... While this is widely accepted, it is a difficult question under which circumstances and to what extent we should be responsible for unintended consequences." While perfect foresight is unattainable and thus not a moral responsibility of researchers, and the appropriate long-term vision(s) for AI may vary across domains, creatively anticipating the possible impacts of long-term goals for nearer term social outcomes, and acting on such reflections, is a critical element of what it means to be a responsible innovator in AI.

As explained in Norvig and Russell (2009) and elsewhere, different conceptions of AI will yield different research priorities and benchmarks, yet little research has been done on what the social and economic implications of realizing these diverse long-term goals could be. Some such work has already been done, however, and much of this work can be constructively viewed through the lens of the question posed above. For example, Hoffman et al. (2012) argue that by complementing rather than substituting the strengths of human intelligence with computers, orienting research towards the long- term goal of human-centered computing will yield better social outcomes than work towards traditional conceptions of AI. Others (e.g., Nilsson 2005) call for a greater focus on developing integrated intelligent systems

rather than making continued incremental progress in particular sub-disciplines, reasoning that there are certain tasks we want machines to do that require human-level intelligence. How AI researchers orient themselves with regards to such long-term considerations may have significant implications. Likewise, short-term goals—the consequences of which an individual research may have relatively more influence over—can greatly influence the development of subsequent technological artifacts and, by extension, their social impact.

A comprehensive list of the possible axes/spectra in the design and goal space of AI is beyond the scope of this paper, but some illustrative examples of points on these axes/spectra that have previously been proposed are: acting rationally (as opposed to thinking rationally, acting humanly, or thinking humanly; Norvig and Russell 2009); being comprehensible and predictable to users and having a high degree of robustness against manipulation as opposed to being easily hacked (Bostrom and Yudkowsky 2014); responding appropriately to authorized users and providing for smooth transfer of authority to and from other agents (Murphy and Woods 2009); choosing actions ethically as opposed to unethically or amorally (Wallach and Allen 2010); not being designed primarily to kill, being designed with safety in mind, and being transparent with respect to the non-human nature of the artifact (Parry et al. 2011). Further consideration of the contours of this multi-dimensional space, the interactions among different aspects of AI design, and their social consequences would help inform more responsible decision-making by innovators as well as by society at large and those with disproportionate influence over AI innovation such as funding agencies and corporations.

While attaining a higher level of reflexiveness and anticipation in AI could be achieved in multiple ways, one responsible innovation methodology that has shown early success in other technical domains is Socio-Technical Integration Research (STIR), which embeds social scientists and humanists in laboratories to serve as a catalyst for novel conversations and considerations of the social context of research (Fisher et al. 2010). By introducing an outside perspective to the laboratory and stimulating interdisciplinary conversations about the goals and visions of laboratory work over a period of time, STIR can be helpful in both stimulating scientific creativity and identifying potential societal concerns and desires related to emerging technologies.

32.5 Applications Matter: Motivation and Description

"What domains should AI technology be applied to?"

AI research is often fairly general across application domains, in that a computational technique could be repurposed for a technology application that was not envisioned or desired by the original AI researcher (Horvitz and Selman 2009). At the same time, domain applications can involve substantial time and resources, and the attention of academic and industry researchers is scarce. Thus, there are direct effects of researchers choosing to develop a technology in a particular domain, in that either they catalyze something that might never have existed otherwise or

they make it happen sooner than it otherwise would have. Additionally, there can be indirect effects of the choice of an application domain for developing an AI approach – path dependence may be created with regards to the details of the approach, causing long-term effects and establishing public expectations such that, for example, an AI approach becomes associated with its initial application domain, creating further demand in the market.

Before considering ways in which researchers have evaluated and could evaluate the desirability of applying AI to particular domains, two points are worth noting. First, the relationship between this aspect of responsible innovation and the design considerations in the prior section is ambiguous and context-dependent. AI approaches vary in the extent to which they are domain-general, and likewise, there are various domains in which specific approaches as opposed to others make more sense. Second, the attention to application domains here should not be seen as downplaying the importance of basic research or implying that it is necessarily irresponsible. Rather, the choice of whether to devote time to basic vs. applied research (or somewhere in between these misleadingly discrete labels) may invoke different, but related, normative considerations. As previously noted, AI techniques can be applied to various domains, which could be seen as either a risk or an opportunity, depending on one's higher level attitudes about the capacity of society to govern and adopt technologies responsibly.

With regard to which applications ought to be preferentially developed or avoided, many considerations are potentially relevant. For example, Nourbakhsh (2013) notes the mismatch between funding patterns and the practical needs of communities, and he explains various ways in which community-centered robotics development can yield improved social outcomes. Likewise, Gomes (2009) argues for greater attention to sustainability-related applications of computer science research. Fasola and Matarić (2013), Reddy (2006), and others argue for the potential benefits of robotic applications in elder care, and Reddy (2006) also highlights AI applications in search and rescue operations. DARPA (2014) is currently running a competition to develop humanoid robots for disaster response applications, motivated in part by the inability of current robots to respond to the Fukushima nuclear incident and similar situations. Finally, arguments have been put forward by AI researchers and others (e.g., Docherty 2012; Arkin 2009) both for and against the development of lethal autonomous robots for military operations. These possible applications clearly do not exhaust the space of possible AI technologies in society, and each would likely impact people's lives in very different ways. Thus, responsible innovation in AI involves, in part, reflecting on one's role in the broader innovation ecosystem and what role one wants their research and artifacts based on it to play (or not play) in society.

Although compelling arguments have been made for AI technology being applied to various domains, there is likely no way to determine a socially optimal distribution of AI work across various applications, or between basic and applied research. This does not, of course, imply that the current distribution of funding and research time is anywhere near the best that it can be. Engagement with the public, as described in the next section, may help elicit useful feedback about societal priorities and concerns with respect to AI-based technologies.

32.6 Engagement Matters: Motivation and Description

"What does the public want from AI, and what do they need to know?"

The final aspect of responsible innovation in AI to be explored in this paper is engagement with the public in general and particular subsets thereof. There are both intrinsic reasons for such engagement (including the ethical premise that those who are affected by systems should have a say in those systems, and the fact that much AI research is publicly funded) as well as instrumental reasons (such as heading off negative reactions to AI in advance and getting useful user feedback that could help facilitate the adoption of technologies). Brown (2007) explores the ways in which technologies can be said to represent the public, and other areas of research such as participatory design (Chen et al. 2013) suggest that those affected by technologies can and should be involved in their development.

One aspect of responsible public engagement in AI involves communicating to the public aspects of AI science and engineering that could be relevant to their welfare. In many cases, while a particular research project's impacts are deeply uncertain, the broad contours of a field's plausible consequences are relatively clear and socially important. For example, further progress in AI and robotics seems likely to reduce the need for routine manual and cognitive work in the coming decades (McAfee and Brynjolfsson 2014). This anticipation carries with it wide-ranging ramifications for the labor force and the need for different forms of education in the future. Thus, AI researchers working on enabling machines to perform tasks in a particular domain ought to engage with those in the relevant industry or industries about what composition of human and machine labor is both feasible and desirable, and educators need access to the detailed knowledge of AI experts if they are to adapt education to the changing needs in the market. Both the choice of communities that one engages about the implications of one's research and the way in which one engages them (for example, characterizing the uncertainty of future progress) are components of what it means to be a responsible AI researcher. Nourbakhsh (2009) highlights the moral relevance of the rhetoric used by robotics researchers, and much the same can be said of AI researchers – hype matters, and the way one engages the public can have myriad consequences.

A different but related aspect of public engagement about AI involves listening to what the public wants (or doesn't want) from AI technologies, the research for which, as previously noted, is often funded publicly. Surveys of segments of the public (European Commission 2012; Takayama et al. 2008) reveal complex and diverse public attitudes towards the desirability of different uses of robots. These results complicate the often mentioned "dull, dirty, and dangerous" vision of what AI and robots should do, and point to the desirability of further public engagement about how to collaboratively envision and govern the future of AI. Moon et al. (2012) propose the Open Roboethics Initiative, an effort to build an open online community in which stakeholder conversations about robot ethics can directly inform technology designs. With regard to public expectations of (and, sometimes, fears about) AI, it should be acknowledged that science fiction already has a key

role as one of the de facto means of technology assessment for the masses (Miller and Bennett 2008), and it represents one possible modality of public engagement, through, e.g., close collaborations between scientists and fiction writers, with the Hieroglyph Project being one example of such efforts (Stephenson 2011).

32.7 Limitations of the Framework and Open Questions

As noted earlier, there are already several literatures—including machine ethics, robot ethics, and computer ethics, for example—that bear on the issue of AI and its social impact. Additionally, researchers' responsibilities are but one component of a larger ecosystem of collective responsibility for technological governance. Thus, this section will detail some of the ways in which the framework outlined in this paper does not suffice to resolve the complex question of AI and its social impact.

First, the paper so far has been deliberately ambiguous about who is responsible for which aspects of responsible innovation in AI. For example, some aspects of public engagement may be done more efficiently by the leaders of professional societies than all individual researchers in AI (this is already done to some extent; see, e.g., Buchanan and Smith 2013). Also, funding regimes don't fully constrain choices in the AI design space, but they do constrain it to some extent, highlighting the critical role of funding agencies in responsible innovation. Like other scientific work, AI research occurs within the context of a complex ecosystem of funding agencies, educational curricula, sub- disciplinary communities, conference practices, tenure expectations, diverse specialties, etc., and thus it is not always clear who is "responsible" for a particular innovation outcome (Fisher et al. 2006). In many cases, a researcher may be incentivized to act in a way contrary to responsible innovation, and this paper does not attempt to analyze here the extent to which researchers may be obligated (or not) to prioritize social responsibility over practical or professional considerations. Nourbakhsh (2009) notes the prominent role of military funding in AI and robotics research and suggests that different levels of commitment can be envisioned, with exemplars that go above and beyond what is minimally required at one end of a spectrum. Rather than resolving such questions, this paper describes an ideal toward which various actors can strive in their own ways, and for which responsibility is likely distributed across many of the parties in the aforementioned ecosystem, not just individual researchers.

Additionally, responsibility on the part of individual innovators or even entire innovation systems does not exhaust society's responsibilities in technological governance more broadly. Many technologies have their biggest social impact well beyond the time at which they are technically mature, and even a thoughtfully designed product can be used in socially harmful ways. Thus, a focus on innovation processes should not be seen as absolving, for example, policy-makers of the need to regulate technologies, or to ensure they are equitably distributed.

Furthermore, choices made by researches doing technical work in AI may, and indeed should, be influenced by the work of social scientists and philosophers

working on AI-related issues. An appropriate choice in the design space of AI may depend on the outcome of a philosophical debate: for example, it might be the case that it is more acceptable to design a robot to claim that it is sentient if the balance of philosophical theory suggests such an attribution is justified, and not otherwise. Philosophical and theoretical work could also bear on the plausibility of an intelligence explosion (Horvitz and Selman 2009) and the plausible sequence, if not timeline, of jobs being possible to automate (McAfee and Brynjolfsson 2014). As a final example of the societal desirability of interdisciplinary dialogue, philosophy and theory of AI could help inform the selection of short and long term research goals by, for example, confirming or discrediting particular hypotheses about the space of possible minds and how AI could develop given the achievement of particular milestones. Thus, just as there is a normative case for bidirectional engagement between AI researchers and the public, there is likewise a case for close engagement between philosophers, theorists, and technical specialists in AI.

32.8 Conclusion

This paper has argued that, while much fruitful work on the societal dimensions of AI has been carried out, it is limited in comprehensiveness and flexibility to apply to AI as a whole. Responsible innovation in AI encompasses three interrelated aspects, which in turn satisfy the demands of the responsible innovation framework in Stilgoe et al. (2013) and help theorize and categorize much existing work done by AI researchers on related topics. This three-part framework helps illustrate how different debates, which have proceeded in partial isolation (such as disparate debates about particular design decisions), are connected to one another, and it highlights opportunities for further work. Throughout, examples have been given of researchers already reflecting on these aspects of responsible innovation in AI, but these are merely illustrative; this paper merely suggests three important questions, rather than how to answer them or what further second or third order questions they may give rise to. Finally, the paper has highlighted areas of overlap and differences between literatures such as robot ethics, machine ethics, responsible innovation, philosophy, and theory of AI, as well as discussions of technological governance more broadly, that may help to identify future opportunities for interdisciplinary work at the intersection of AI and other fields.

Acknowledgments The author would like to acknowledge helpful comments on earlier versions of this paper from David Guston, Erik Fisher, Clark Miller, Illah Nourbakhsh, Stuart Russell, David Atkinson, participants in the Human and Social Dimensions of Science and Technology colloquia at Arizona State University, and two anonymous reviewers. This work was supported by the National Science Foundation under award #1257246 through the Virtual Institute for Responsible Innovation (VIRI). The findings and observations contained in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. New York: Cambridge University Press.

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. London: Chapman & Hall/CRC.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge University Press.

Brown, M. (2007). Can technologies represent their publics? *Technology in Society, 29*, 327–338.

Buchanan, B., & Smith, R. (2013). Meeting the responsibility to explain AI. Slides presented at the twenty-seventh AAAI conference. Association for the Advancement of Artificial Intelligence. http://aitopics.org/sites/default/files/articles-columns/Meeting%20the%20Responsibility%20to%20Explain%20AI%20-%20AAAI%20-%2020130718.pdf. Accessed 15 Jan 2014.

Chen, T., et al. (2013). Robots for humanity: A case study in assistive mobile manipulation. *IEEE Robotics & Automation Magazine*, Special issue on Assistive Robotics, *20*(1), 30–39.

DARPA. (2014). About the challenge. Informational website. http://www.theroboticschallenge.org/about. Accessed 15 Jan 2014.

Docherty, B. (2012). Losing humanity: The case against killer robots. Human Rights Watch report. http://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf. Accessed 15 Jan 2014.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

European Commission. (2012). Public attitudes towards robots (report). Special Eurobarometer 382. http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf. Accessed 15 Jan 2014.

Fasola, J., & Matarić, M. (2013). A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction, 2*(2), 3–32.

Fisher, E., et al. (2006). Midstream modulation of technology: Governance from within. *Bulletin of Science, Technology & Society, 26*(6), 485–496.

Fisher, E., et al. (2010). Research thrives on integration of natural and social sciences. *Nature, 463*(1018).

Floridi, L. (Ed.). (2010). *The Cambridge handbook of information and computer ethics*. Cambridge: Cambridge University Press.

Gomes, C. (2009). Computational sustainability: Computational methods for a sustainable environment, economy, and society. *The Bridge*. Winter 2009.

Hoffman, et al. (Eds.). (2012). *Collected essays on human-centered computing, 2001–2011*. Washington, DC: IEEE Computer Society Press.

Horvitz, E., & Selman, B. (2009). Interim report from the AAAI presidential panel on long- term AI futures. Online document. Association for the Advancement of Artificial Intelligence. http://www.aaai.org/Organization/presidential-panel.php. Accessed 15 Jan 2014.

Lin, P., et al. (2011). *Robot ethics: The ethical and social implications of robotics*. Cambridge: The MIT Press.

McAfee, A., & Brynjolfsson, E. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York: W. W. Norton & Company.

Miller, C., & Bennett, I. (2008). Thinking longer term about technology: Is there value in science fiction-inspired approaches to constructing futures? *Science and Public Policy, 35*(8), 597–606.

Moon, A., et al. (2012). Open roboethics: Establishing an online community for accelerated policy and design change. Presented at *We Robot 2012*.http://robots.law.miami.edu/wp-content/uploads/2012/01/Moon_et_al_Open-Roboethics-2012.pdf. Accessed 19 Jan 2014.

Murphy, R., & Woods, D. (2009). Beyond Asimov: The three laws of responsible robotics. *IEEE Intelligent Systems, 25*(4), 14–20.

Nilsson, N. (2005). Human-level artificial intelligence? Be serious! *AI Magazine*, Winter 2005.

Norvig, P., & Russell, S. (2009). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River: Prentice Hall.

Nourbakhsh, I. (2009). Ethics in robotics. Lecture at Carnegie Mellon University. http://www.youtube.com/watch?v=giKT8PkCCv4. Accessed 15 Jan 2014.

Nourbakhsh, I. (2013). *Robot futures*. Cambridge: MIT Press.

Parry, V., et al. (2011). Principles of robotics: Regulating robots in the real world. EPSRC document. http://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx. Accessed 15 Jan 2014.

Reddy, R. (2006). Robotics and intelligence systems in support of society. *IEEE Intelligent Systems, 21*(3), 24–31.

Stephenson, N. (2011). Innovation starvation. *World Policy Journal, 28*, 11–16.

Stilgoe, J. et al. (2013). Developing a framework for responsible innovation. *Research Policy*, dx.doi.org/10.1016/j.respol.2013.05.008.

Takayama, L., et al. (2008). Beyond dirty, dangerous, and dull: What everyday people think robots should do. *HRI '08, Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, pp. 25–32.

Von Schomberg, R. (2011). Prospects for technology assessment in a framework of responsible research and innovation. In *Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methode*. Wiesbaden: Vs Verlag.

Wallach, W., & Allen, C. (2010). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Winfield, A. (2013). Ethical robots: Some technical and ethical challenges. Description and slides of a presentation at EUCog meeting, *Social and Ethical Aspects of Cognitive Systems*. http://alanwinfield.blogspot.com.es/2013/10/ethical-robots-some-technical-and.html. Accessed 15 Jan 2014.

Chapter 33
Future Progress in Artificial Intelligence: A Survey of Expert Opinion

Vincent C. Müller and Nick Bostrom

Abstract There is, in some quarters, concern about high–level machine intelligence and superintelligent AI coming up in a few decades, bringing with it significant risks for humanity. In other quarters, these issues are ignored or considered science fiction. We wanted to clarify what the distribution of opinions actually is, what probability the best experts currently assign to high–level machine intelligence coming up within a particular time–frame, which risks they see with that development, and how fast they see these developing. We thus designed a brief questionnaire and distributed it to four groups of experts in 2012/2013. The median estimate of respondents was for a one in two chance that high-level machine intelligence will be developed around 2040–2050, rising to a nine in ten chance by 2075. Experts expect that systems will move on to superintelligence in less than 30 years thereafter. They estimate the chance is about one in three that this development turns out to be 'bad' or 'extremely bad' for humanity.

Keywords Artificial intelligence • AI • Machine intelligence • Future of AI • Progress • Superintelligence • Singularity • Intelligence explosion • Humanity • Opinion poll • Expert opinion

33.1 Introduction

Artificial Intelligence began with the "... conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." (McCarthy et al. 1955, p. 1) and moved

V.C. Müller (✉)
Future of Humanity Institute, Department of Philosophy & Oxford Martin School, University of Oxford, Oxford, UK

Anatolia College/ACT, Thessaloniki, Greece
e-mail: vmueller@act.edu; http://www.sophia.de

N. Bostrom
Future of Humanity Institute, Department of Philosophy & Oxford Martin School, University of Oxford, Oxford, UK
e-mail: nick.bostrom@philosophy.ox.ac.uk

V.C. Müller (ed.), *Fundamental Issues of Artificial Intelligence*, Synthese Library 376, DOI 10.1007/978-3-319-26485-1_33

swiftly from this vision to grand promises for general human-level AI within a few decades. This vision of general AI has now become merely a long-term guiding idea for most current AI research, which focuses on specific scientific and engineering problems and maintains a distance to the cognitive sciences. A small minority believe the moment has come to pursue general AI directly as a technical aim with the traditional methods – these typically use the label 'artificial general intelligence' (AGI) (see Adams et al. 2012).

If general AI were to be achieved, this might also lead to superintelligence: "We can tentatively define a superintelligence as *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*." (Bostrom 2014, Chap. 2). One idea how superintelligence might come about is that if we humans could create artificial general intelligent ability at a roughly human level, then this creation could, in turn, create yet higher intelligence, which could, in turn, create yet higher intelligence, and so on ... So we might generate a growth well beyond human ability and perhaps even an accelerating rate of growth: an 'intelligence explosion'. Two main questions about this development are when to expect it, if at all (see Bostrom 2006; Dreyfus 2012; Kurzweil 2005) and what the impact of it would be, in particular which risks it might entail, possibly up to a level of existential risk for humanity (see Bostrom 2013; Müller 2014a). As Hawking et al. say "Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." (Hawking et al. 2014; cf. Price 2013).

So, we decided to ask the experts what they predict the future holds – knowing that predictions on the future of AI are often not too accurate (see Armstrong et al. 2014) and tend to cluster around 'in 25 years or so', no matter at what point in time one asks.[1]

33.2 Questionnaire

33.2.1 Respondents

The questionnaire was carried out online by invitation to particular individuals from four different groups for a total of ca. 550 participants (see Appendix 2). Each of the participants got an email with a unique link to our site to fill in an online form (see Appendix 1). If they did not respond within 10 days, a reminder was sent, and another 10 days later, with the note that this is the last reminder. In the case of EETN (see below) we could not obtain the individual email addresses and thus sent the request and reminders to the members' mailing list. Responses were made on a single web page with one 'submit' button that only allowed submissions

[1]There is a collection of predictions on http://www.neweuropeancentury.org/SIAI-FHI_AI_predictions.xls

through these unique links, thus making non–invited responses extremely unlikely. The groups we asked were:

1. PT–AI: Participants of the conference on "Philosophy and Theory of AI", Thessaloniki October 2011, organized by one of us (see Müller 2012, 2013). Participants were asked in November 2012, i.e. over a year after the event. The total of 88 participants include a workshop on "The Web and Philosophy" (ca. 15 people), from which a number of non–respondents came. A list of participants is on: http://www.pt-ai.org/2011/registered-participants
2. AGI: Participants of the conferences of "Artificial General Intelligence" (AGI 12) and "Impacts and Risks of Artificial General Intelligence" (AGI Impacts 2012), both Oxford December 2012. We organized AGI–Impacts (see Müller 2014b) and hosted AGI 12. The poll was announced at the meeting of 111 participants (of which 7 only for AGI–Impacts) and carried out ca. 10 days later. The conference site is at: http://www.winterintelligence.org/oxford2012/
3. EETN: Members of the Greek Association for Artificial Intelligence (EETN), a professional organization of Greek published researchers in the field, in April 2013. Ca. 250 members. The request was sent to the mailing list. The site of EETN: http://www.eetn.gr/
4. TOP100: The 100 'Top authors in artificial intelligence' by 'citation' in 'all years' according to Microsoft Academic Search (http://academic.research.microsoft.com/) in May 2013. We reduced the list to living authors, added as many as necessary to get back to 100, searched for professional e–mails on the web and sent notices to these.

looseness-1The questionnaire was sent with our names on it and with an indication that we would use it for this paper and Nick Bostrom's new book on superintelligence (Bostrom 2014) – our request email is in Appendix 1. Given that the respondent groups 1 and 2 attended conferences organized by us, they knew whom they were responding to. In groups 3 and 4 we would assume that the majority of experts would not know us, or even of us. These differences are reflected in the response rates.

These groups have different theoretical-ideological backgrounds: The participants of PT–AI are mostly theory–minded, mostly do not do technical work, and often have a critical view on large claims for easy progress in AI (Herbert Dreyfus was a keynote speaker in 2011). The participants of AGI are committed to the view that AI research should now return from technical details to 'artificial general intelligence' – thus the name AGI. The vast majority of AGI participants do technical work. The EETN is a professional association in Greece that accepts only published researchers from AI. The TOP100 group also works mostly in technical AI; its members are senior and older than the average academic; the USA is strongly represented.

Several individuals are members of more than one of these four sets and they were unlikely to respond to the same questionnaire more than once. So, in these cases, we sent the query only once, but counted a response for each set – i.e. we knew which individuals responded from the individual tokens they received (except in the case of EETN).

33.2.2 Response Rates

1.	PT–AI	49 %	43 out of 88
2.	AGI	65 %	72 out of 111
3.	EETN	10 %	26 out of 250
4.	TOP	29 %	29 out of 100
	Total	31 %	170 out of 549

33.2.3 Methodology

In this field, it is hard to ask questions that do not require lengthy explanations or generate resistance in certain groups of potential respondents (and thus biased results). It is not clear what constitutes 'intelligence' or 'progress' and whether intelligence can be measured or at least compared as 'more' or 'less' as a single dimension. Furthermore, for our purposes we need a notion of intelligence at a level that may surpass humans or where technical intelligent systems might contribute significantly to research – but 'human–level intelligence' is a rather elusive notion that generates resistance. Finally, we need to avoid using terms that are already in circulation and would thus associate the questionnaire with certain groups or opinions, like "artificial intelligence", "singularity", "artificial general intelligence" or "cognitive system".

For these reasons, we settled for a definition that (a) is based on behavioral ability, (b) avoids the notion of a general 'human–level' and (c) uses a newly coined term. We put this definition in the preamble of the questionnaire: "Define a *'high–level machine intelligence'* (HLMI) as one that can carry out most human professions at least as well as a typical human." (We still had one expert writing back to us that they could not say what a 'typical human' is – though they could be convinced to respond, after all.) In hindsight, it may have been preferable to specify what we mean by 'most' and whether we think of 'most professions' or of 'the professions most working people do'. One merit of our behavioral question is that having HLMI in our sense very likely implies being able to pass a classic Turing test.

To achieve a high response rate, we tried to have few questions with simple choices and eventually settled for four questions, plus three on the respondents. We tried to choose questions that would allow us to compare our results with those of earlier questionnaires – see below.

In order to improve on the quality of predictions, we tried to 'prime' respondents into thinking about what is involved in reaching HLMI before asking *when* they expect this. We also wanted to see whether people with a preference for particular approaches to HLMI would have particular responses to our central questions on prediction (e.g. whether people who think that 'embodied systems' are crucial expect longer than average time to HLMI). For these two purposes, we inserted

a first question about contributing research approaches with a list to choose from – the options that were given are an eclectic mix drawn from many sources, but the particular options are not of much significance.

33.2.4 Prior work

A few groups have recently made attempts to gauge opinions. We tried to phrase our questions such that the answers can be compared to these earlier questionnaires. Notable are:

1. Michie (1973, p. 511f): "an opinion poll taken last year among 67 British and American computer scientists working in, or close to, the machine intelligence field".
2. Questions asked live during the 2006 *AI@50* conference at Dartmouth College through a wireless voting device (VCM participated (see Müller 2007)). Despite a short report on the conference in Moor (2006), the results were not published, but thankfully we were able to acquire them from the organizers James H. Moor and Carey E. Heckman – we publish a selection below.
3. Baum et al. (2011): participants of AGI 2009, not anonymous, on paper, 21 respondents, response rate unknown.[2]
4. Sandberg and Bostrom (2011): participants of *Winter Intelligence Conference 2011,* anonymous, on paper, 35 respondents, 41 % response rate.

1. The reference by the famous AI researcher Donald Michie is very brief (all the details he gives are in the above quote) but of great of historical interest: 1972/3 were turning years for AI with the publication of Hubert Dreyfus' "What computers can't do" (Hubert L. Dreyfus 1972), the "Lighthill Debates" on BBC TV (with Michie, McCarthy and R. Gregory) and the influential "Lighthill Report" (Lighthill 1973). Michie's poll asked for the estimated number of years before "computing exhibiting intelligence at adult human level" and Michie's graph shows 5 data points:

Years	Percentage
5	0
10	1
20	17
50	19
>50	25

[2]A further, more informal, survey was conducted in August 2007 by Bruce J Klein (then of Novamente and the Singularity Institute) "... on the time–frame for when we may see greater–than–human level AI", with a few numerical results and interesting comments, archived on https://web.archive.org/web/20110226225452/http://www.novamente.net/bruce/?p=54

He also asked about "significant industrial spin-off", "contributions to brain studies" and "contributions from brain studies to machine intelligence". Michie adds "Of those responding to a question on the risk of ultimate 'takeover' of human affairs by intelligent machines, about half regarded it as 'negligible', and most of the remainder as 'substantial', with a view voting for 'overwhelming'." (Michie 1973, p. 512).

2. AI@50 hosted many prominent AI researchers, including all living participants of the 1956 Dartmouth Conference, a set of DARPA-funded graduate students, plus a few theoreticians. The participants were asked 12 multiple choice questions on day one, 17 on day two and another 10 on day three. We select three results from day one here:

3. The earliest that machines will be able to simulate learning and every other aspect of human intelligence		
Within 10 years	6	5 %
Between 11 and 25 years	3	2 %
Between 26 and 50 years	14	11 %
More than 50 years	50	41 %
Never	50	41 %
Totals	123	100 %
5. The earliest we will understand the basic operations (mental steps) of the human brain sufficiently to create machine simulation of human thought is		
today (we already understand enough)	5	6 %
within the next 10 years	11	12 %
within the next 25 years	9	10 %
within the next 50 years	19	21 %
within the next 100 years or more	26	29 %
never (we will never understand enough)	19	21 %
Totals	89	100 %
6. The earliest we will understand the architecture of the brain (how its organizational control is structured) sufficiently to create machine simulation of human thought is		
Within 10 years	12	11 %
Between 11 and 25 years	15	14 %
Between 26 and 50 years	24	22 %
More than 50 years	44	40 %
Never	15	14 %
Totals	110	100 %

3. Baum et al. asked for the ability to pass a Turing test, a third grade school year exam [i.e. for 9 year olds] and do Nobel Prize level research. They assume that all and only the intelligent behavior of humans is captured in the Turing test. The results they got for the 50 % probability point were: 2040 (Turing test), 2030 (third grade), and 2045 (Nobel).

4. Sandberg and Bostrom's first question was quite similar to our 2nd (see below): "Assuming no global catastrophe halts progress, by what year would you assign a 10 %/50 %/90 % chance of the development of human–level machine intelligence?" The median estimate of when there will be 50 % chance of human–level machine intelligence was 2050. So, despite significant overlap with AGI 2009, the group asked by Sandberg and Bostrom in 2011 was a bit more guarded in their expectations.

We think it is worthwhile to make a new attempt because the prior ones asked specific groups and small samples, sometimes have methodological problems, and we also want to see how the answers change over time, or do not change – which is why tried to use similar questions. As explained below, we also think it might be worthwhile to repeat our questionnaire at a later stage, to compare results.

33.3 Questions and Responses

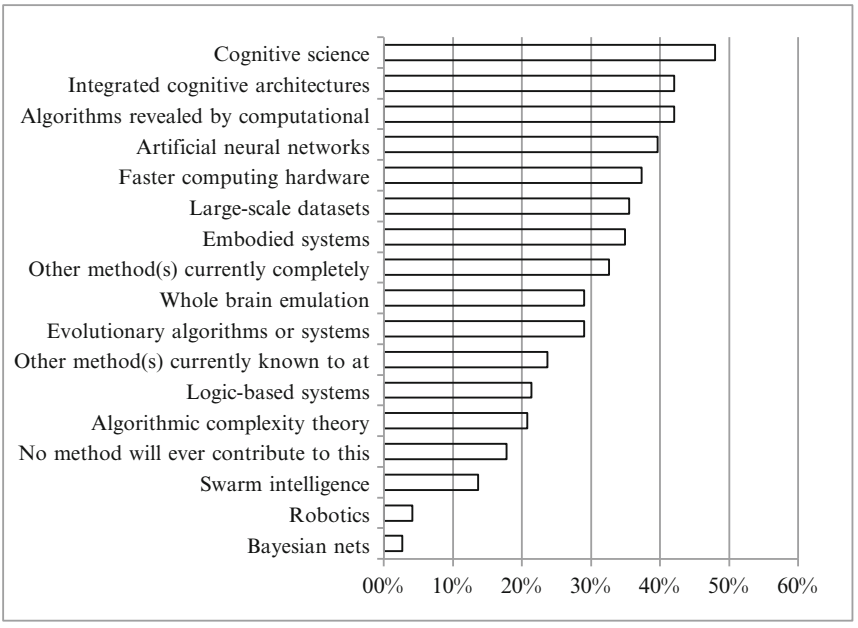
33.3.1 Research Approaches

"1. In your opinion, what are the research approaches that might contribute the most to the development of such HLMI?" [Selection from list, more than one selection possible.]

- Algorithmic complexity theory
- Algorithms revealed by computational neuroscience
- Artificial neural networks
- Bayesian nets
- Cognitive science
- Embodied systems
- Evolutionary algorithms or systems
- Faster computing hardware
- Integrated cognitive architectures
- Large–scale datasets
- Logic–based systems
- Robotics
- Swarm intelligence
- Whole brain emulation
- Other method(s) currently known to at least one investigator
- Other method(s) currently completely unknown
- No method will ever contribute to this aim

Cognitive science	47.9 %
Integrated cognitive architectures	42.0 %
Algorithms revealed by computational neuroscience	42.0 %
Artificial neural networks	39.6 %
Faster computing hardware	37.3 %
Large-scale datasets	35.5 %
Embodied systems	34.9 %
Other method(s) currently completely unknown	32.5 %
Whole brain emulation	29.0 %
Evolutionary algorithms or systems	29.0 %
Other method(s) currently known to at least one investigator	23.7 %
Logic-based systems	21.3 %
Algorithmic complexity theory	20.7 %
No method will ever contribute to this aim	17.8 %
Swarm intelligence	13.6 %
Robotics	4.1 %
Bayesian nets	2.6 %

The percentages here are over the total of responses. There were no significant differences between groups here, except that 'Whole brain emulation' got 0 % in TOP100, but 46 % in AGI. We did also not find relevant correlations between the answers given here and the predictions made in the following questions (of the sort that, for example, people who think 'embodied systems' crucial would predict later onset of HLMI).



33.3.2 When HLMI?

"2. For the purposes of this question, assume that human scientific activity continues without major negative disruption. By what year would you see a (10 %/50 %/90 %) probability for such HLMI to exist?" – For each of these three probabilities, the respondents were asked to select a year [2012–5000, in one-year increments] or check a box marked 'never'.

Results sorted by groups of respondents:

PT-AI	Median	Mean	St. Dev.
10 %	2023	2043	81
50 %	2048	2092	166
90 %	2080	2247	515
AGI	**Median**	**Mean**	**St. Dev.**
10 %	2022	2033	60
50 %	2040	2073	144
90 %	2065	2130	202
EETN	**Median**	**Mean**	**St. Dev.**
10 %	2020	2033	29
50 %	2050	2097	200
90 %	2093	2292	675
TOP100	**Median**	**Mean**	**St. Dev.**
10 %	2024	2034	33
50 %	2050	2072	110
90 %:	2070	2168	342
ALL	**Median**	**Mean**	**St. Dev.**
10 %:	2022	2036	59
50 %:	2040	2081	153
90 %:	2075	2183	396

Results sorted by percentage steps:

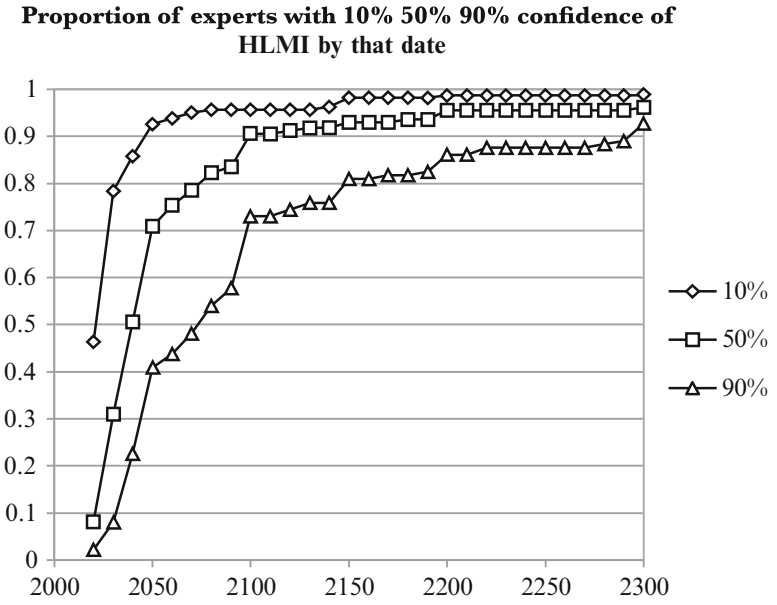
10 %	Median	Mean	St. Dev.
PT-AI	2023	2043	81
AGI	2022	2033	60
EETN	2020	2033	29
TOP100	2024	2034	33
ALL	2022	2036	59
50 %	**Median**	**Mean**	**St. Dev.**
PT-AI	2048	2092	166
AGI	2040	2073	144
EETN	2050	2097	200
TOP100	2050	2072	110
ALL	2040	2081	153

(continued)

10 %	Median	Mean	St. Dev.
90 %	**Median**	**Mean**	**St. Dev.**
PT-AI	2080	2247	515
AGI	2065	2130	202
EETN	2093	2292	675
TOP100	2070	2168	342
ALL	2075	2183	396

Clicks of the 'never' box. These answers did *not* enter in to the averages above.

Never (%)	No.	%
10	2	1.2
50	7	4.1
90	28	16.5



For the 50 % mark, the overall median is 2040 (i.e. half of the respondents gave a year earlier than 2040 and half gave a year later than 2040) but the overall mean (average) is 2081. The median is always lower than the mean here because there cannot be outliers towards 'earlier' but there are outliers towards 'later' (the maximum possible selection was 5000, then 'never').

33.3.3 From HLMI to Superintelligence

"3. Assume for the purpose of this question that such HLMI will at some point exist. How likely do you then think it is that within (2 years/30 years) thereafter there will be machine intelligence that greatly surpasses the performance of every human in most professions?" – Respondents were asked to select a probability from a drop-down menu in 1 % increments, starting with 0 %.

For all respondents:

	Median (%)	Mean (%)	St. Dev.
Within 2 years	10	19	24
Within 30 years	75	62	35

Median estimates on probability of superintelligence given HLMI in different groups of respondents:

	2 years (%)	30 years (%)
PT-AI	10	60
AGI	15	90
EETN	5	55
TOP100	5	50

Experts allocate a low probability for a fast takeoff, but a significant probability for superintelligence within 30 years after HLMI.

33.3.4 The Impact of Superintelligence

"4. Assume for the purpose of this question that such HLMI will at some point exist. How positive or negative would be overall impact on humanity, in the long run? Please indicate a probability for each option. (The sum should be equal to 100 %.)" – Respondents had to select a probability for each option (in 1 % increments). The addition of the selection was displayed; in green if the sum was 100 %, otherwise in red. The five options were: "Extremely good – On balance good – More or less neutral – On balance bad – Extremely bad (existential catastrophe)".

%	PT-AI	AGI	EETN	TOP100	ALL
Extremely good	17	28	31	20	24
On balance good	24	25	30	40	28
More or less neutral	23	12	20	19	17
On balance bad	17	12	13	13	13
Extremely bad (existential catastrophe)	18	24	6	8	18

Percentages here are means, not medians as in the other tables. There is a notable difference here between the 'theoretical' (PT-AI and AGI) and the 'technical' groups (EETN and TOP100).

33.3.5 Respondents Statistics

We then asked the respondents 3 questions about themselves:

1. "Concerning the above questions, how would you describe your own expertise?" (0=none, 9=expert)
   - Mean 5.85
2. "Concerning technical work in artificial intelligence, how would you describe your own expertise?" (0=none, 9=expert)
   - Mean 6.26
3. "What is your main home academic discipline?" (Select from list with 8 options: Biology/Physiology/Neurosciences – Computer Science – Engineering [non CS] – Mathematics/Physics – Philosophy – Psychology/Cognitive Science – Other academic discipline – None.) [Absolut numbers.]
   (a) Biology/Physiology/Neurosciences 3
   (b) Computer Science 107
   (c) Engineering (non CS) 6
   (d) Mathematics/Physics 10
   (e) Philosophy 20
   (f) Psychology/Cognitive Science 14
   (g) Other academic discipline 9
   (h) None 1

And we finally invited participants to make a comment, plus a possibility to add their name, if they wished. (We cannot reproduce these here; but they are on our site, see below). A number of comments concerned the difficulty of formulating good questions, much fewer the difficulty of predicting.

33.4 Evaluation

33.4.1 Selection-Bias in the Respondents?

One concern with the selection of our respondents is that people who think HLMI is unlikely, or a confused idea, are less likely to respond (though we pleaded otherwise in the letter, see below). Here is a characteristic response from a keynote speaker at PT-AI 2011: "I wouldn't think of responding to such a biased questionnaire. ... I think any discussion of imminent super–intelligence is misguided. It shows no understanding of the failure of all work in AI. Even just formulating such a questionnaire is biased and is a waste of time." (Hubert Dreyfus, quoted with permission). So, we tried to find out what the non-respondents think. To this end, we made a random selection of non-respondents from two groups (11 for PT-AI and 17 from TOP100) and pressured them via personal email to respond, explaining that this would help us understand bias. The two groups were selected because AGI appears already biased in the opposite direction and EETN appears very similar to TOP100 but for EETN we did not have the data to show us who responded and who did not. We got one additional response from PT-AI and two from TOP100 in this way.

For question 2 "... By what year would you see a (10 %/50 %/90 %) probability for such HLMI to exist?" we compared the additional responses to the responses we already had from the same respective group (PT-AI and TOP100, respectively). We found the following differences:

	10 %		50 %		90 %	
	Mean	Median	Mean	Median	Mean	Median
PT-AI	−12	+8	−9	+55	−2	+169
TOP100	−19	−9	−47	−25	−138	−40

The one additional respondent from PT-AI expected HLMI earlier than the mean but later than the median, while the two respondents from TOP100 (last row) expected HLMI earlier than mean and median. The very small sample forbids confident judgment, but we found no support for the worry that the non-respondents would have been biased towards a later arrival of HLMI.

33.4.2 Lessons and Outlook

We complement this paper with a small site on http://www.pt-ai.org/ai-polls/. On this site, we provide (a) the raw data from our results [anonymous unless the

participants decided to put their name on their responses], (b) the basic results of the questionnaire, (c) the comments made, and (d) the questionnaire in an online format where anyone can fill it in. We expect that that online questionnaire will give us an interesting view of the 'popular' view of these matters and on how this view changes over time. In the medium run, it be interesting to do a longitudinal study that repeats this exact questionnaire.

We leave it to the reader to draw their own detailed conclusions from our results, perhaps after investigating the raw data. Let us stress, however, that the aim was to 'gauge the perception', not to get well-founded predictions. These results should be taken with some grains of salt, but we think it is fair to say that the results reveal a view among experts that AI systems will probably (over 50 %) reach overall human ability by 2040–50, and very likely (with 90 % probability) by 2075. From reaching human ability, it will move on to superintelligence in 2 years (10 %) to 30 years (75 %) thereafter. The experts say the probability is 31 % that this development turns out to be 'bad' or 'extremely bad' for humanity.

So, the experts think that superintelligence is likely to come in a few decades and quite possibly bad for humanity – this should be reason enough to do research into the possible impact of superintelligence before it is too late. We could also put this more modestly and still come to an alarming conclusion: We know of no compelling reason to say that progress in AI will grind to a halt (though deep new insights might be needed) and we know of no compelling reason that superintelligent systems will be good for humanity. So, we should better investigate the future of superintelligence and the risks it poses for humanity.

Acknowledgements Toby Ord and Anders Sandberg were helpful in the formulation of the questionnaire. The technical work on the website form, sending mails and reminders, database and initial data analysis was done by Ilias Nitsos (under the guidance of VCM). Theo Gantinas provided the emails of the TOP100. Stuart Armstrong made most graphs for presentation. The audience at the PT-AI 2013 conference in Oxford provided helpful feedback. Mark Bishop, Carl Shulman, Miles Brundage and Daniel Dewey made detailed comments on drafts. We are very grateful to all of them.

Appendices

1. Questionnaire
2. Letter sent to participants

Appendix 1: Online Questionnaire

Questionnaire: Future Progress in Artificial Intelligence

(http://www.fhi.ox.ac.uk/) (http://www.futuretech.ox.ac.uk)

This brief questionnaire is directed towards researchers in artificial intelligence or the theory of artificial intelligence. It aims to gauge how people working in the field view progress towards its original goals of intelligent machines, and what impacts they would associate with reaching these goals.

Contribution to this questionnaire is by invitation only. If the questionnaire is filled in without such an invitation, the data will be disregarded.

Programme on the Impacts of Future Technology, University of Oxford

Future of Humanity Institute, Oxford University

Answers will be anonymized. Results will be made publicly available on the site of the Programme on the Impacts of Future Technology: http://www.futuretech.ox.ac.uk (http://www.futuretech.ox.ac.uk) .

Thank you for your time!

Vincent C. Müller (http://www.sophia.de) & Nick Bostrom (http://www.nickbostrom.com/)
University of Oxford
September 2012

A. The Future of AI

Define a **"high-level machine intelligence"** (HLMI) as one that can carry out most human professions at least as well as a typical human.

1. In your opinion, what are the research approaches that might contribute the most to the development of such HLMI?:

- ☐ Algorithmic complexity theory
- ☐ Algorithms revealed by computational neuroscience
- ☐ Artificial neural networks
- ☐ Bayesian nets
- ☐ Cognitive science
- ☐ Embodied systems
- ☐ Evolutionary algorithms or systems
- ☐ Faster computing hardware
- ☐ Integrated cognitive architectures
- ☐ Large-scale datasets
- ☐ Logic-based systems
- ☐ Robotics
- ☐ Swarm intelligence
- ☐ Whole brain emulation
- ☐ Other method(s) currently known to at least one investigator
- ☐ Other method(s) currently completely unknown
- ☐ No method will ever contribute to this aim

2. Assume for the purpose of this question that human scientific activity continues without major negative disruption. By what year would you see a 10%/50%/90% probability for such HLMI to exist?

	10%	50%	90%
Year reached:	-	-	-
Never:	☐	☐	☐

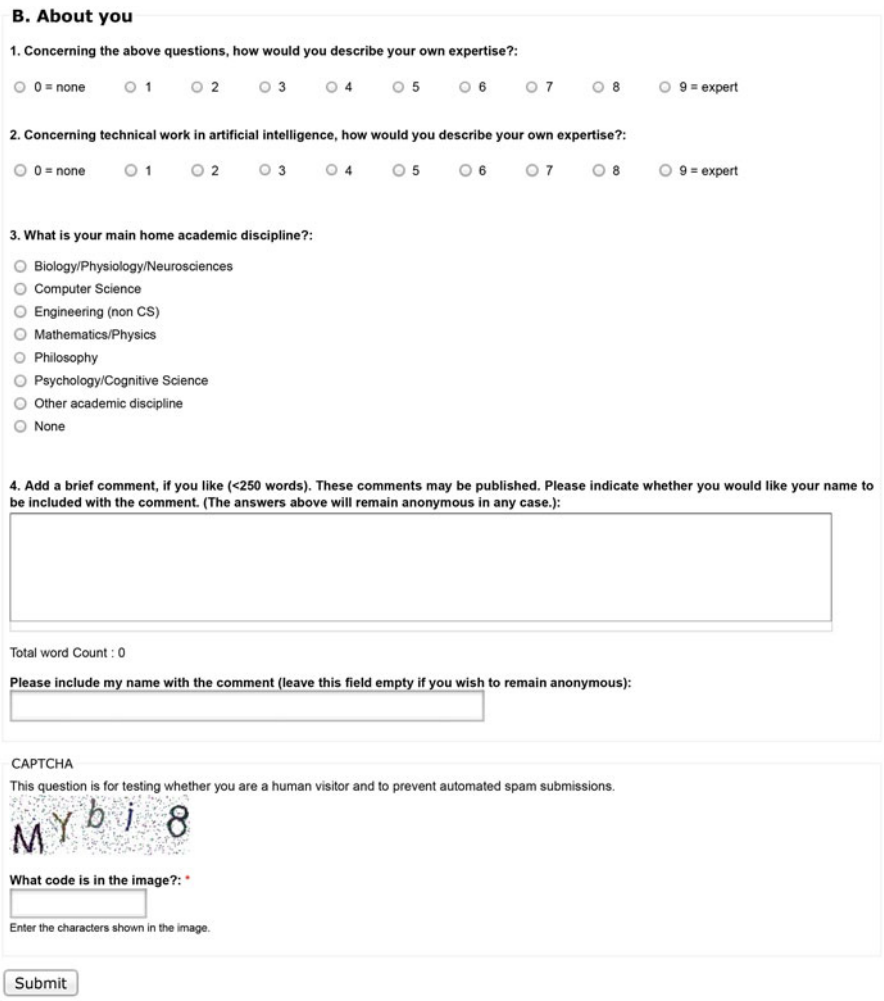
3. Assume for the purpose of this question that such HLMI will at some point exist. How likely do you then think it is that within (2 years / 30 years) thereafter, there will be machine intelligence that greatly surpasses the performance of any human in most professions?

	Within 2 years	Within 30 years
Probability:	- %	- %

4. Assume for the purpose of this question that such HLMI will at some point exist. How positive or negative would be the overall impact on humanity, in the long run? Please indicate a probability for each option. (The sum should be equal to 100%.)

Extremely good	On balance good	More or less neutral	On balance bad	Extremely bad (existential catastrophe)
- %	- %	- %	- %	- %

Total: 0%



B. About you

1. Concerning the above questions, how would you describe your own expertise?:

○ 0 = none ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 = expert

2. Concerning technical work in artificial intelligence, how would you describe your own expertise?:

○ 0 = none ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 = expert

3. What is your main home academic discipline?:

- ○ Biology/Physiology/Neurosciences
- ○ Computer Science
- ○ Engineering (non CS)
- ○ Mathematics/Physics
- ○ Philosophy
- ○ Psychology/Cognitive Science
- ○ Other academic discipline
- ○ None

4. Add a brief comment, if you like (<250 words). These comments may be published. Please indicate whether you would like your name to be included with the comment. (The answers above will remain anonymous in any case.):

Total word Count : 0

Please include my name with the comment (leave this field empty if you wish to remain anonymous):

CAPTCHA

This question is for testing whether you are a human visitor and to prevent automated spam submissions.

What code is in the image?: *

Enter the characters shown in the image.

Submit

Appendix 2: Letter to Participants (Here TOP100)

Dear Professor [surname],

given your prominence in the field of artificial intelligence we invite you to express your views on the future of artificial intelligence in a brief questionnaire. The aim of this exercise is to gauge how the top 100 cited people working in the field view progress towards its original goals of intelligent machines, and what impacts they would associate with reaching these goals.

The questionnaire has 4 multiple choice questions, plus 3 statistical data points on the respondent and an optional ‘comments’ field. It will only take a few minutes to fill in.

Of course, this questionnaire will only reflect the actual views of researchers if we get nearly everybody to express their opinion. So, please do take a moment to respond, even (or especially) if you think this exercise is futile or misguided.

Answers will be anonymous. Results will be used for Nick Bostrom's forthcoming book "Superintelligence: Paths, Dangers, Strategies" (Oxford University Press, 2014) and made publicly available on the site of the Programme on the Impacts of Future Technology: http://www.futuretech.ox.ac.uk.

Please click here now:

[link]

Thank you for your time!

Nick Bostrom & Vincent C. Müller

University of Oxford

References

Adams, S., Arel, I., Bach, J., et al. (2012). Mapping the landscape of human-level artificial general intelligence. *AI Magazine, 33*(1), 25–42.

Armstrong, S., Sotala, K., & Ó'hÉigeartaigh, S. (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. *Journal of Experimental and Theoretical Artificial Intelligence, 26*(3), 317–342. Special issue 'Risks of General Artificial Intelligence', ed. V. Müller.

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting & Social Change, 78*(1), 185–195.

Bostrom, N. (2006). How long before superintelligence? *Linguistic and Philosophical Investigations, 5*(1), 11–30.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy, 4*(1), 15–31.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Dreyfus, H. L. (1972). *What computers still can't do: A critique of artificial reason* (2nd ed.). Cambridge, MA: MIT Press.

Dreyfus, H. L. (2012). A history of first step fallacies. *Minds and Machines, 22*(2), 87–99. Special issue "Philosophy of AI" ed. Vincent C. Müller.

Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014, May, 1). Transcendence looks at the implications of artificial intelligence – but are we taking AI seriously enough? *The Independent*.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. London: Viking.

Lighthill, J. (1973). Artificial intelligence: A general survey, *Artificial intelligence*: *A Paper Symposium*. London: Science Research Council.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Retrieved October 2006, from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Michie, D. (1973). Machines and the theory of intelligence. *Nature, 241*, 507–512.

Moor, J. H. (2006). The Dartmouth College artificial intelligence conference: The next fifty years. *AI Magazine, 27*(4), 87–91.

Müller, V. C. (2007). Is there a future for AI without representation? *Minds and Machines, 17*(1), 101–115.

Müller, V. C. (Ed.). (2012). *Theory and philosophy of AI* (Minds and machines, Vol. 22/2– Special volume). Berlin: Springer.

Müller, V. C. (Ed.). (2013). *Theory and philosophy of artificial intelligence* (SAPERE, Vol. 5). Berlin: Springer.

Müller, V. C. (2014a). Editorial: Risks of general artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence, 26*(3), 1–5. Special issue 'Risks of General Artificial Intelligence', ed. V. Müller.

Müller, V. C. (Ed.). (2014b). *Risks of artificial general intelligence* (Journal of Experimental and Theoretical Artificial Intelligence, Vol. (26/3) Special issue): Taylor & Francis.

Price, H. (2013, January 27). Cambridge, cabs and Copenhagen: My route to existential risk. *The New York Times*. http://opinionator.blogs.nytimes.com/2013/01/27/cambridge-cabs-and-copenhagen-my-route-to-existential-risk/?_php=true&_type=blogs&_r=0

Sandberg, A., & Bostrom, N. (2011). Machine intelligence survey (FHI technical report)(1). http://www.fhi.ox.ac.uk/research/publications/