Philosophical Studies Series

Don Berkich
Matteo Vincenzo d'Alfonso *Editors*

On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence

Themes from IACAP 2016

Springer

Philosophical Studies Series

Volume 134

More information about this series at http://www.springer.com/series/6459

Don Berkich • Matteo Vincenzo d'Alfonso
Editors

On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence

Themes from IACAP 2016

Springer

Editors
Don Berkich
Department of Humanities
Texas A&M University Corpus Christi
Corpus Christi, TX, USA

Matteo Vincenzo d'Alfonso
Dipartimento di Studi Umanistici
Università di Ferrara
Ferrara, Italy

ISSN 0921-8599 ISSN 2542-8349 (electronic)
Philosophical Studies Series
ISBN 978-3-030-01799-6 ISBN 978-3-030-01800-9 (eBook)
https://doi.org/10.1007/978-3-030-01800-9

Library of Congress Control Number: 2018965206

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction to This Volume

Don Berkich

Abstract The 2016 meeting of the International Association for Computing and Philosophy brought together a highly interdisciplinary consortium of scholars eager to share their current research at the increasingly important intersection of a number of fields, including computer science, robotics engineering, artificial intelligence, logic, biology, cognitive science, economics, sociology, and philosophy. This introductory chapter serves to organize and connect the discussions across these domains of inquiry, describing both the insights and broad relevance of the research well represented in this volume.

Keywords International Association for Computing and Philosophy · Computation · Information · Logic · Epistemology · Science · Cognition · Mind · Robotics · Ethics · Trust · Privacy · Justice

1.1 The International Association for Computing and Philosophy

The intersection between philosophy and computing is curiously expansive, as the articles in this volume amply demonstrate. New vistas of inquiry are being discovered and explored in a way that neither field alone, philosophy nor computer science, would suggest. An exemplar for the fruitfulness of interdisciplinary work can be found in the International Association for Computing and Philosophy (IACAP).

At its inception in the mid-1980s, Computing and Philosophy conferences were almost wholly devoted to discussing the pedagogical uses of the freshly-deployed desktop computer. Much of that work now seems quaint in light of the many

D. Berkich (✉)
Department of Humanities, Texas A&M University – Corpus Christi, Corpus Christi, TX, USA
e-mail: berkich@gmail.com

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_1

ways in which computers and networks have subsequently become integral to the functioning of the modern university's educational mission. Yet it is interesting to note that philosophers were 'early adopters', front and center in discussions of how best to adapt the computer to help in teaching.

This pedagogical focus would persist through the early 1990s. However, the long history tying philosophical, mathematical, and computational investigations together (the work of Hobbes and Leibniz looms particularly large in this regard) would soon draw philosophical and mathematical logicians, computer scientists, neuro-scientists, ethicists, roboticists, psychologists, information theorists, and philosophers of mind into discussions at turns historical, foundational, and applicable. In the subsequent decades, individual threads of inquiry and subsequent discussions have been woven into the fabric of important research agendas well furthered by papers presented at the 2016 meeting of the International Association for Computing and Philosophy at the University of Ferrara, Ferrara, Italy, June 14–17. Hosted by Professors Marcello DAgostino and (my co-editor) Matteo DAlfonso, IACAP 2016 was graciously sponsored by the University of Ferrara, the Department of Economics and Management, and by the Dipartimento di Studi Umanistici.

The 21 contributions to this volume neatly represent a cross section of 40 papers, 4 keynote addresses, and 8 symposia as they cut across fully six distinct research agendas. Now, I take it that an editor's duty is not merely to describe the ways in which the contributions further the research agendas, but to help frame and better set those agendas for readers and researchers alike. Briefly, then, this volume begins with foundational studies in (1) Computation and Information, (2) Logic, and (3) Epistemology and Science. Research into computational aspects of (4) Cognition and Mind leads neatly into (5) Moral Dimensions of Human-Machine Interaction, followed finally by broader social and political investigations into (6) Trust, Privacy, and Justice. Consider each in turn.

1.2 Computation and Information

In the abstract it is conventional to formally characterize computation in various extensionally equivalent ways – viz., Turing Machine Computability (Turing 1936), $\lambda-$Calculability (Church 1932), Primitive Recursion (Gödel 1931), or even Abacus Computability (Boolos and Jeffrey 1989). Such purely formal characterizations themselves do little, however, to answer a host of important, subtly difficult, and deeply related questions: What is computation? Does computation elucidate mechanism, or does mechanism elucidate computation? Do computational processes describe a natural kind, or is virtually any physical process at some level of description computational in nature? Likewise, what is information? What is the relationship between information, on the one hand, and computation, on the other? Are some or all physical processes inherently informational, or is the notion of information simply a conceptual scheme by which physical processes may be interpreted? Notice that all of these questions, and many other related questions

besides, are foundational in precisely the sense in which answers to them are presupposed by such questions as, is the brain a kind of information processing organ? Indeed, this last question motivates the investigations taken up in our first two contributions.

Following Piccinini's excellent survey of the problem (Piccinini 2017), we want to know which complex physical systems implement computations (artifacts like smart-phones, say, or naturally occurring systems like mammalian nervous systems) formally characterized by some conception of algorithm or other and which do not (a freshly painted wall, a stone garden pathway, or a pile of sand). A chunk of carved quartz crystal would not be a smart-phone, no matter how careful the carving and close the resemblance, presumably because the quartz crystal lacks the smart-phone's capacity to, variously, realize, concretize, or implement computations as defined formally and abstractly. Piccinini dubs this 'the problem of concrete computation'.

If concrete computation implements formal computation merely by it happening to be the case that there exists a state-preserving mapping from formal computational states into physical states (Putnam 1975), then *pancomputationalism* threatens – to borrow Piccinini's terminology. That is, *any* sufficiently complex physical system – among them the molecules of paint drying on a wall, a pile of sand, or our quartz crystal smart-phone facsimile – will implement formal computation. Thus to the question of whether the brain is computational in nature, we answer, only in the same vacuous sense in which any physical system is computational in nature. Put another way, concrete computation is eliminated as a natural kind on such an account.

Paul Schweizer's "Computation in Physical Systems: A Normative Mapping Account" and Vincenzo Fano, et al.'s "When is a Computation Realized in a Concrete Physical System?", both informed by Piccinini's sketch of the terrain and his terms of the debate, offer competing analyses of concrete computation in attempting to counter pancomputationalism.

Explicitly echoing Dennett's Intentional Stance, whereby intentional states like beliefs and desires are ascribed as such insofar as doing so yields explanation and prediction (Dennett 1987), Schweizer proposes that, while there is no concrete computational natural kind – and, thus, any physical process can in principle implement formal computations – we avoid the threat pancomputationalism appears to pose to computationalist theories of mind by virtue of the fact that some physical processes are more suited to our pragmatic interests in concrete computation. Taking the *computational stance*, the physical properties of the smart-phone (in terms of high and low voltages and the complex electrical properties of semiconductors) make it vastly more useful to us than any attempt at treating its quartz crystal analog computationally would do, despite the fact that we could in principle take Schweizer's computational stance with respect to it. Likewise the brain: With respect to explanation and prediction, it is more useful to take the computational stance with respect to neural processes, so computational theories of mind are not undermined by the fact that we could also treat piles of sand as concrete computations.

Where Dennett hedges somewhat on outright rejection of the existence of intentional states (he calls himself a 'quasi-realist' with respect to them), Schweizer sees the computationalist stance as justifying, explicitly, anti-realism with respect to concrete computation. It is not altogether clear, though, whether the possibility that concrete computation boils down to an observer-relative computational stance licenses thorough-going anti-realism about concrete computational systems. After all, as Dennett himself points out (Dennett 1987) in regards to original intentionality, there must be *some* states of the organism which serve to ground the success of taking the intentional stance in explanation and prediction. Similarly, if taking the computational stance is successful in explaining and predicting the behavior of some physical system, surely the most one can assert is a kind of agnosticism with respect to concrete computation.

That taking the computational stance is sometimes useful in explanation and prediction and sometimes not seems itself a curious fact to be explained. An anti-realist would of course point out that this is question-begging: As Schweizer well argues, our pragmatic interests in taking one stance or another just are the whole of the explanation. Nothing more need be added. Nevertheless, the agnostic may suspect that *some* concrete property of the physical system computationally viewed suits it for computational explanation and prediction.

Oberholzer and Gruner seek to resolve a long-standing debate between Floridi and Fetzer concerning the nature of information. For Floridi – and, for the many reasons Floridi has presented (2007) – information must be true to count as such. Although Oberholzer and Gruner are quick to point out that Floridi allows for information to be both factive and instructional, the grist for the Floridi/Fetzer debate is on information in its factive sense. Oberholzer and Gruner take the factive sense to imply that it is more than merely representative, significant, or faithful: It is necessarily *propositional* given that it (when not instructional) must be true. This, they point out, is out-of-step with classical views of information whereby it (i) is conceived as data structured in such a way as to be communicable and thus usable, whether true or false, (ii) runs afoul of arguments begun by Fetzer (2014) and vigorously pursued by Scarantino and Piccinini (2010), and, writing as computer scientists themselves, (iii) raises puzzles for how best to conceive of the stimuli-response relationship involved in engineering behavioral robotic systems.

While their sympathies are clearly on the Fetzer, Scarantino, and Piccinini side of the debate, Oberholzer and Gruner seek a resolution by arguing that it rests on an equivocation over 'information'. Drawing on Frege's distinction between sense and reference, they suggest in light of various arguments by Scarantino and Piccinini that 'information' classically construed is to be understood in terms of the thought a proposition expresses, whereas Floridi's more restrictive, factive notion of information is better suited to the true proposition's reference. Of course, as the Frege Argument (1980) shows, all true propositions have the same reference, so it is not clear whether the Fregean distinction does much useful work here. Nevertheless, by adapting the *Afrikaans* terms 'inligting' and 'informasie' to 'enlightation' and 'information', respectively, Oberholzer and Gruner seek to expose what they

see as the underlying equivocation by applying 'enlightation' to Floridi's more restrictive sense of 'information' so as to preserve the traditional sense and use of 'information'.

While their distinction between enlightation and information may not settle the Floridi/Fetzer debate, Oberholzer and Gruner remind us that the intersection of computing and philosophy imposes a crucial *computational obligato* on philosophical inquiry. That is, computationally informed philosophical inquiry is also constrained by the computational (logical, mathematical) facts and the way those facts bear on engineering questions, modestly echoing the attempt by logical empiricists to impose an *empirical obligato* on philosophical inquiry. Philosophical disputes on matters of computation and information particularly are constrained by the computational facts and fruitful to the extent that they inform simultaneously the technical and the technological. In a theme woven throughout this volume, the computational obligato serves to undermine, if not entirely eliminate, philosophical flights of fancy which may otherwise be thought to impugn philosophical inquiry.

1.3 Logic

As with philosophical inquiry into the nature of computation and information, so too is there much to consider at the intersection of computation and logic. From automatic theorem provers to the development of non-standard logics and even to the history and theory of computation itself, logicians both mathematical and philosophical have laid the foundation for computation and worked alongside engineers to develop and refine computer technology. Two contributions to this volume, Elohim's "Modal $\Omega-$Logic: Automata, Neo-Logicism, and Set-theoretic Realism" and Mario Piazza and Marco Pedicini's "What Arrow's Information Paradox Says (To Philosophers)" are rooted in this tradition, yet each in their own way seek broader philosophical implications.

In perhaps the most technically demanding paper of this volume, Elohim sets out to explicate $\Omega-$logic validity in modal $\Omega-$logic for ZFC set theory. With an account in hand, he uses the flexible notions of coalgebraic logics and automata to come at the same concept from these two other directions. Given the modal nature of $\Omega-$logical validity so-defined, Elohim pauses to describe its epistemic variation and briefly argue that it has application in virtue of its automata definition to computational theories of mind.

Be that as it may, Elohim's quarry here is to draw two lessons from the coalgebraic definition of $\Omega-$logic validity for the philosophy of mathematics. First, Elohim suggests that insofar as $\Omega-$logical validity is purely logical, albeit having modal properties, it justifies neo-logicism in that the conceptual truths of mathematics (at least, insofar as ZFC Set Theoretic truths are concerned) are not stronger or more questionable than the underlying $\Omega-$logic expressing those truths. Second, in a series of arguments, Elohim makes the case that our very grasp of the concept of (the hierarchy of) sets is itself modal in nature to the extent that

we understand the meaning qua intension of the concept. Put more simply, modal Ω−logic supplies the resources we require to flesh-out the intuitive notion of a set meant to be captured by the usual extensional characterization of ZFC set theory but which seems in many ways to exceed it. If so, Elohim reasons, then we *also* have an argument for mathematical platonism. After all, there is some *thing* – the hierarchy of sets, namely – we grasp in our intuitive understanding.

Piazza and Pedicini likewise seek to draw broader philosophical implications – epistemic, in their case – from Arrow's information paradox (Arrow 1962). The context here is economic. Shopping for a new bicycle, I need to have as much information about it as possible to determine whether it is worth the cost. Where the product in question is information itself, I need to have as much information about the information to determine whether it is worth the cost. Yet that is just to have the information, without paying any cost whatsoever. The very act of determining the value of information obliterates its value. So either information cannot be construed as a product in the first place – anathema to a free market capitalist system – or there needs to be state intervention in the free market – also anathema to the free market – by establishing and enforcing intellectual property rights. Piazza and Pedicini point out that the economic literature using Arrow's paradox to justify intellectual property rights assumes I'm faithless: Once I have the information to determine its value, I drop the exchange having got what I set out to obtain in the first place without any cost to me.

To take considerations of intellectual property rights off the table, assume I'm honest. Then, Piazza and Pedicini argue, I nevertheless find myself in an epistemic paradox which echoes the *Meno* paradox: I am either blindly pursuing information, not knowing what it is I pursue, or, already knowing it, have no need to pursue it. Modeling my pursuit of information on Shannon's (cryptographic) Information Theory (Shannon 1949) offers a way to conceptualize certification or verification of the information without its transmission, thus finding a third alternative to blind or unnecessary pursuit of information. That approach aside, Piazza and Pedicini suggest that the larger lesson to be drawn is that curiosity – for the honest agent, at least – comes inevitably at a cost.

1.4 Epistemology and Science

As computational resources are more cheaply deployed in empirical inquiry than ever before, whether it be for the sake of collecting, storing, and analyzing vast troves of data or constructing and verifying computational models of complex systems, epistemic questions arise which challenge received views about the very nature of scientific inquiry.

Carnap neatly summarizes the crucial transition in science from the teleological (Aristotelian) 'why'-questions characteristic of science prior to Newtonian mechanics to their abandonment in favor of the 'how'-questions characteristic of current science:

> In the nineteenth century, certain Germanic physicists, such as Gustav Kirchhoff and Ernst Mach, said that science should not ask "Why?" but "How?" They meant that science should not look for unknown metaphysical agents that are responsible for certain events, but should only describe such events in terms of laws. This prohibition against asking "Why?" must be understood in its historical setting. The background was the German philosophical atmosphere of the time, which was dominated by idealism in the tradition of Fichte, Schelling, and Hegel. These men felt that a description of how the world behaved was not enough. They wanted a fuller understanding, which they believed could be obtained only by finding metaphysical causes that were behind phenomena and not accessible to scientific method. Physicists reacted to this point of view by saying: "Leave us alone with your why-questions. There is no answer beyond that given by the empirical laws." They objected to why-questions because they were usually metaphysical questions.
>
> Today the philosophical atmosphere has changed. In Germany there are a few philosophers still working in the idealist tradition, but in England and the United States it has practically disappeared. As a result, we are no longer worried by why-questions. We do not have to say, "Don't ask why", because now, when someone asks why, we assume that he means it in a scientific, nonmetaphysical sense. He is simply asking us to explain something by placing it in a framework of empirical laws. (Carnap 1998, p. 678)

Carnap thus aligns the shift from why to how-questions with Hempel's covering law model of explanation and prediction (Hempel 1998a), wherein the explanandum is deduced from an explanans containing laws and statements of conditions. The laws in question, whether interpreted as carrying the necessity of causal law or epistemic regularities (Hempel 1998b), explain and predict by deductively justifying the explanadum. Whether a given such deduction serves predictive or explanatory purposes has nothing to do with the deduction per se and everything to do with the scientist's interests.

However crude this gloss surely is, it suffices to highlight the remarkable shift the expanding use of computational methods and so-called Big Data in science have caused. For just as Carnap invites us to drop why-questions and focus exclusively on usefully answerable how-questions, the sophisticated statistical analysis of massive data sets can identify strong correlations without explanatory bearing. Perhaps, then, we should drop how-questions given the size and complexity of the data-sets in favor of *that-questions*: that events are highly correlated regardless of how they are so related confers predictive power without the unnecessary epistemic burden of explanation.

Indeed, our hand may be forced. More sophisticated algorithms for analyzing large data sets for structure beyond mere correlations which might be employed in the service of answering how-questions themselves confront fundamental complexity constraints on what is feasibly computable. The hard limits of those constraints threaten to render large data sets explanatorily impenetrable.

In "Antimodularity: Pragmatic Consequences of Computational Complexity in Scientific Explanation", Luca Rivelli shows how the limits on what is computable in light of complexity constraints for the large input characteristic of scientific inquiry into large systems – e.g., meteorology, ecology, biology, or neurology – raises serious challenges for the epistemic goal of scientific explanation. Specifically, the received view on explanations of such systems is that the system's global behavior can be functionally decomposed into the interactions of sub-systems or

modules whose (simpler) functional features contribute to, and account for, the super-system's features. Thus the modular specification of a complex system – our ability, that is, to describe it in modular terms – is essential to explanation, or so Rivelli argues.

Yet drawing on the example of network analysis by computational means reveals that the general problem of modular specification is at best a matter of approximation given complexity constraints on such computations, limitations quickly discovered even for not especially large systems. The upshot, Rivelli suggests, is that some systems may be expected to be of such a scale that we can have no confidence whatsoever in any modular specification given by algorithm. Rivelli dubs this *antimodularity*, whereby a system exceeds the limits of even approximate specification. Such a system, Rivelli warns, is functionally impenetrable and inexplicable insofar as explanation presupposes some sort of modular specification. Far from simply aiding in the pursuit of scientific explanation, complexity constraints on computational analysis reveal the limits of explanation and, in the example of antimodularity, the possibility of the inexplicable.

If the modular analysis of complex systems can be foiled by the apparently hard limits of computational complexity, perhaps a creative enterprise like computer science can further illuminate the problems encountered in pursuit of reductive explanations. In "A Software-Inspired Constructive View of Nature", Russ Abbott argues that the practice of computer science – that is, constructing novel functional properties by piecing together simpler functional elements in novel ways – *constructive creativity* in Abbott's terms – provides a metaphor which can usefully be applied to better grasp the limits and nature of scientific explanation.

Although Abbott takes care to point out that it is no more than an analogy, the parallels he draws between explanation in computer science and science generally are striking, particularly regarding complex systems and reductive explanation.

The computer scientist has at her disposal a raft of libraries and low-level function calls suited to creatively constructing, building-block fashion, new, more complex functionality. Although an explanation of the resulting functionality can be given in principle at the level of machine-code and register calls, it would be useless so far as the computer scientist's interests over constructed functional capabilities are concerned. Of course, the computer scientist has the luxury of designing the low-level functionality in such a way that it *permits* creative construction.

Are there parallels to creative construction in nature? That is, are physical scientists in roughly the same position as computer scientists in their prospects for giving reductive explanations? If so, then we ought to find parallels to the functional compatibility we employ by design in computer science to achieve creative construction. Abbott points to three physical analogs which, without the convenience of having been designed, nevertheless underwrite natural creative construction: negative interaction energy, or the attractive sub-atomic forces binding particles together in atomic structures; autopoiesis, the oft-criticized notion of self-sustaining and replicating structures; and, altogether specific to the biological, evolution itself. Just how far the analogy between the physical and the computational can be pressed is open to question, but it does recognize that there is presumably

a point of diminishing explanatory relevance the more basic or fundamental the reduction. Just as the computer scientist is properly concerned with features of the available libraries and not the particular states of the microprocessor's registers, the biologist is properly concerned with the organism's capacities and requirements in light of the structure and function of its constitutive organs, say, and not the properties of the sub-atomic particles they contain.

The implications of big data and its computational analysis are no less important for the prospects of explanation in the social and political sciences than the physical sciences, as Teresa Numerico explores in her cautionary "Politics and Epistemology of Big Data: A Critical Assessment". The *exemplum primi* motivating her analysis is Facebook's now infamous 2015 emotional contagion research on nearly a million of its users which, the authors concluded, showed that positive and negative emotions propagate in social networks (Kramer 2014). Setting aside the obvious concerns Numerico raises regarding the issue of informed consent which the Facebook researchers utterly ignored, she rightly points out that the wealth of data we create in our online activities should disallow the social science researcher from any pretense that anonymity is a sufficient protection for subjects and, furthermore, the data itself can only be analyzed by algorithms which in turn embed biases which should disabuse researchers of faith in them as objective research tools. Numerico argues in particular that machine learning algorithms deployed for the analysis of big data are *epistemically opaque* in the sense that the methods leading to their results cannot be verified by human researchers. Epistemically, machine learning constitutes a kind of black box in the social scientific endeavor. Thus, what can be quantified about individuals' behavior in online environments leads, by the sheer vastness of the data, to analyses which neither respect the individual nor are answerable to human researchers.

Whether pointing to the hard limits (vis-a-vis the complexity constraints Rivelli marks) or the soft limits (the epistemic opacity of learning algorithms Numerico describes) of computational methods in scientific inquiry – or, indeed, whether computer science can help illuminate natural science, as Abbott argues – the reliance of scientific inquiry on computational methods begs for greater attention by scientists and philosophers of science alike on the ways in which those methods are informing and changing our understanding of scientific explanation and prediction.

1.5 Cognition and Mind

Big data and its analysis by computational methods are relatively new techniques in biology, physics, and sociology. Nearly from their respective inceptions, however, cognitive science and computation have been pursued so tightly in tandem as to have sprung from the same philosophical roots. Turing's (1936) demonstration of the existence of the Universal Turing Machine, a turing machine that can compute any of the denumerable functions computable by some turing machine, in conjunction with the Church-Turing thesis (Boolos and Jeffrey 1989) that any function

computable by some effective procedure is turing machine computable, almost immediately raised the intriguing question of whether cognitive capacities, suitably decomposed in terms of underlying cognitive functions, were not themselves turing machine computable. Put another way, we want to know whether the class of cognitive functions is wholly contained in the class of turing machine computable functions, where the effective procedures in question are neurological in nature.

The audacious hypothesis of cognitive science is that cognition itself is explicable in computational terms. As computer technology advances, the hand-in-glove fit of computation and cognition creates richer opportunities for the study of cognition, perception, action, and their artificial counterparts. The five contributions here capture the breadth and depth of some of the resulting research agendas.

Tjostheim and Leister explore the philosophical foundations bearing on the empirical dimensions of the study of telepresence in their "Telepresence and the Role of the Senses". Consider, for a somewhat concrete example, the operator of a remotely operated submersible such as those deployed by marine scientists and in underwater oil exploration. Using two cameras on the ROS permits depth perception for close work, but it costs the operator the disconcerting feeling of being at once on the ship and submerged 200 meters, simultaneously. Tactile and olfactory senses align with being ship-bound, visual senses with being ROS-bound.

Vaguely understood as the feeling of *being there*, telepresence is something video game and virtual reality designers are eager to exploit for entertainment purposes by creating richly detailed environments. One can, for example, explore a virtual Los Angeles. Pointing out that our capacity to experience telepresence can shed light on the nature of the cognition of sense perception, Tjostheim and Leister are particularly interested in the role of *affordance* in telepresence. Although much work, they note, remains to fully flesh-out the notion that objects present properties suited to their usefulness in agency, what one feels one can 'do' with the virtual objects one finds in a virtual environment surely bears on telepresence. Here, of course, the video-game industry is deeply engaged in developing virtual affordances in the service of telepresence and story-telling. Commercial interests aside, however, the philosophical implications of telepresence range from support for the spinozistic proposition that comprehension entails, at least for an instant, belief, to the nature of subjective experience and methodological puzzles of phenomenological surveys. That said, Tjostheim and Leister's research is both preliminary and promising. As they point out, conceptual analyses of affordance and telepresence are largely unsettled and rich in opportunities for further research.

M. Christina Amoretti et al. target conceptual analysis itself in their "Ontologies, Mental Disorders and Prototypes". The logical advantages recommending traditional conceptual analysis by giving individually necessary and jointly sufficient conditions contend with the withering criticisms of the later Wittgenstein and the accumulation of empirical evidence that the role of concepts in cognition is better understood in terms of exemplars or prototypes. Using medical practice with respect to psychological diagnosis as a particularly illuminating example, the authors argue that the typicality conditions used in descriptions of mental disorders found in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V)

offer a treasure-trove for formulating new approaches to concept-mapping and ontology development for application in artificial intelligence.

In this contribution, Amoretti et al. use the Ontology Web Language-Description Logic (OWL-DL) – containing class, role, and individual constructs – to develop a schizophrenia spectrum formal ontology which, better than previous attempts, captures the syndrome, or prototypical description, the DSM-V employs. The point of the exercise is to demonstrate in application the limitations of traditional conceptual analysis in representing knowledge built from the ground up, as it were, not on necessary and sufficient conditions, but on typicality conditions. In representing the DSM-V, however, the authors explain that OWL-DL is a conceptual procrustean bed. Their proposal here, and the direction of their current research, is to adopt a hybrid approach which brings together traditional conceptual analysis (so far as is feasible) with the geometric format of a conceptual spaces analysis. Instances of a concept are modeled as locations in regions (perhaps overlapping) which correspond to concepts. Spatial characteristics like being centered in a convex region can be used to represent the prototypical instance of the corresponding concept, while distances between locations in a region can represent similiarity relations between conceptually related individuals. The promise of such a hybrid approach to knowledge representation would presumably be wider than the diagnosis of mental disorders and apply to conceptual analysis generally.

The application of computational methods to knowledge representation surely has promise in modeling domains of inquiry, yet the more headline-grabbing application is to modeling human neurology, even to large-scale models which seek to explain and predict the behavior of the entire brain. In "Large-scale Computer Models of the Brain: Is There a 'Right' Level of Detail?", Edoardo Datteri takes up the puzzle of just how much detail in brain-modelling is necessary to gain explanatory traction.

A common assumption on the part of the various, flashy whole-brain modeling projects, Datteri points out, is that explanation of human behavior will only be possible with models of exceptionally fine granularity – down to the level of modeling the functional features of individual neurons. Eliasmith and Trujillo (2014), however, argue using the analogy to large-scale climate modeling that there is no right answer to just how fine-grained a model must be: The granularity of the model depends on trade-offs between the questions being asked and the computational resources available. The goal of Datteri's novel and carefully argued contribution is to first (and quickly) dismiss the relevance of the abundance or scarcity of computational resources to the epistemic question of what counts as a sufficient neuroscientific explanation of behavior. He then turns to the difficult task of sussing out how the explanandum dictates the appropriate neuroscientific explanans qua computer modeling, specifically as to just how fine-grained the computer model must be to count as a satisfactory explanation. Put another way, and assuming such explanations involve mechanistic decompositions of complex to simpler neurological mechanisms, how simple must the explaining mechanism (computationally modeled) be to count as a satisfactory explanation?

Datteri's answer is nuanced. In some of the cases Datteri describes, what is to be explained neuroscientifically wholly dictates how course-grained or fine-grained the modeling must be, but in many other cases it does not, contrary to Eliasmith and Trujillo's whole-sale assertion. Where it does not simpliciter, further epistemic principles are required guide modeling efforts – to determine, that is to say, the explanatory adequacy of a given model. What is at stake in these philosophical puzzles is nothing less than determining what counts as a good neuroscientific explanation insofar as those explanations rely on computational methods in modeling neurological systems, as Datteri himself points out.

Computational methods and technology surely have application to modeling in cognitive science and extensive epistemic ramifications, at least for that particular science. That said, information and computation theories more broadly may have implications for long-standing problems in philosophy. d'Alfonso takes up one such problem in his "Virtual Information in the Light of Kant's Practical Reason." Consider the fundamental theorem of deduction,

$$\Gamma \vdash \varphi \text{ iff } \vdash \Gamma \rightarrow \varphi$$

The fact that any set of postulates Γ entailing some theorem φ is equivalent to a tautology is epistemically problematic: The entailment appears to be informative, yet the tautalogy, being necessarily true, carries no information whatsoever. Thus no information is conveyed by the fact that the euclidean postulates entail the pythagorean theorem. If there is no information gained, then there is no epistemic gain, either. Nothing new is learned in the proof of the theorem, since it is already contained in – as it were – the postulates given the tautological equivalence expressed by the fundamental theorem.

D'Agostino and Floridi (2009) propose to rescue the presumed epistemic gain of the entailment by appeal to their concept of *virtual information*. That is, in the course of a natural (as opposed to axiomatic) deduction, temporary assumptions are made and later discharged. These temporary assumptions do briefly convey information and thereby signal epistemic gain in the course of the deduction. As d'Alfonso points out, D'Agostino takes this to be a Kantian solution. Deductions which make recourse to the information carried by temporary assumptions are a priori, as are all deductions, but synthetic as well. An axiomatic deduction which makes no such dischargeable assumptions is, in D'Agostino's scheme, a priori analytic insofar as it conveys no virtual information at all.

d'Alfonso in this contribution seeks to explain, in Kantian terms, the nature of the virtual information in question. In particular, he argues that while the context of the epistemically gainful deduction is *theoretical* in terms of Kant's distinction between theoretical and practical reasoning, our capacity to employ virtual information depends entirely on our practical, or normative, reasoning. Thus the 'should' in the logic professor's exhortation, "You should temporarily assume P so as to infer Q" represents, in Kantian terms, the practical activity essential to the deduction. The epistemic gain of the deduction is in the practical reasoning deployed in its construction, if d'Alfonso is correct.

d'Alfonso draws on the Kantian distinction between practical and theoretical reasoning to develop the D'Agostino and Floridi notion of virtual information, arguing that Kant's distinction neatly explains the epistemic gain from the practical reasoning demanded by mastering an entailment as opposed to its absence on the tautological – and, thus, theoretical – side of the fundamental theorem's equivalence. The long-standing problem in question is an epistemic one raised by the facts of deduction. The solution proposed here hinges on the epistemic relevance of information understood through the lens of Kant's distinction between practical and theoretical reasoning.

Indeed, speculative philosophy has at least since Descartes and the epistemic turn in philosophy focused on the presuppositions the possibility of knowledge (and its character, objects, etc.) place on cognition. What must the mind be like, philosophers have asked, such that knowledge is possible? Competing answers and vigorous debates about the nature of mind have in effect staked out a sort of solution space for cognitive architecture. These proposals, however, have heretofore been speculative – unmoored from any from any sort of verification or testing.

Complicating matters is the metaphysical dimension of the mind-body problem. This may seem an odd claim to make. After all, the mind-body problem, understood as the problem of nature of mind and the nature of body when the properties and relations of mind and body differ so radically as to be utterly distinct, is ordinarily cast first and foremost – if not wholly – as a metaphysical problem. Solutions to the mind-body problem seek to account for the tremendous gap between mental properties and physical properties by working out ways in which the mental and the physical may or may not be distinct substances. For the dualist, the difference in properties signals a difference in substance. The problem then becomes how to account for the apparent ways in which these different substances interact, which generates a plethora of dreaded philosophical 'isms': interactionism, epiphenomenalism, parallelism, etc.

Surely *part* of the problem is metaphysical. Yet the philosophers taking their cue from Descartes, including Leibniz, Berkeley, Hume, Spinoza, Kant, etc., were at least as keen to understand the nature of the properties of mind which constitute the mind-body problem in the first place. Metaphysics aside, here we find extensive investigations which are doxastic, affective, and agential features of mind – cognitive architectures, in short, which, while frequently covered by the mantel of metaphysics, can usefully be divorced from particular metaphysical presuppositions. Neutral monism, for example, does not entail a humean bundle-of-perceptions view of the mind any more than cartesian interactionism excludes it. For the most part, speculations about cognitive architectures can, as psychology has endlessly demonstrated, be conducted while largely ignoring the metaphysics of the mind-body problem.

Yet if investigations of the nature of cognition are to be more than merely speculative, it must be possible to inquire how they would or would not work in practice. Understood in a functionalist sense, the notion of 'work' here opens the door to a sort of computationalist check on what is possible, cognitively speaking.

That is, computation provides a sort of proving ground for philosophical speculation about the nature of cognition.

For instance, the widest gap, that between the empiricist and rationalist traditions in philosophy, is reflected – unintentionally, perhaps – in the gap between deep-learning and logic-based inferential approaches to artificial intelligence. In "A Kantian Cognitive Architecture", Richard Evans finds inspiration in the kantian synthesis of empiricism and rationalism to implement a computational model which builds on the strengths of deep-learning and inferential approaches via a computational counterpart to the kantian synthesis. Though admittedly nascent, Evans' project shows promise on the various tests Evans applies to its current implementation. His approach is encouraging inasmuch as it shows how unsupervised learning can be used on a paucity of data points to more efficiently interpret and systematize the data. As Evans points out, the capacity for such efficient deep learning can in general be obtained by building in domain-specific prior constraints. The trouble is how to build in prior constraints for efficient deep learning which are not domain-specific. Enter Kant. Evans takes Kant's Analytic of Principles to provide a set of general prior constraints, shows how they can be rendered in logical terms which can then be translated into computationally tractable terms, and tests the resulting implementation.

There is much for philosophers and computer scientists alike to glean from Evans' project. In the former case, Evans strives to hover as close as possible to Kant's statements of the principles; his logical analysis of those principles is alone a significant contribution. In the later case, his computational implementation of the logical analysis shows how a synthetic approach to developing general prior constraints on deep learning for the sake of demonstrably improved efficiency can be derived in a principled, yet not domain-specific, way. First to last, Evans' ambitious efforts are a step at making good on promises of the philosophical relevance of computation to philosophy. Of course, Evans' project invites a great deal of further discussion on both the philosophical and computational sides. Yet that is rather the point: The specific moves Evans makes on matters of kantian interpretation, logical rendering, and computational implementation each open broad spaces for further debate, discussion, and collaboration.

The need for collaboration between philosophers, neuroscientists, and computer scientists for our grasp of cognition and the development of a cognitive science is where many epistemic questions will, perhaps, find answers. At the same time, the ongoing rapid development of computer technology has raised at least as many moral and legal normative questions which have drawn the attention of a large share of ethicists, roboticists, and researchers in artificial intelligence. Some of the questions, as we shall shortly see, are rather specific, pointing out pitfalls of implementation approaches that ought ethically be avoided, while others are quite general, raising questions about the very nature of a society which is increasingly characterized by the interactions between human beings and the machines they create.

1.6 Moral Dimensions of Human-Machine Interaction

In "Machine Learning and Irresponsible Inference: Morally Assessing the Training Data for Image Recognition Systems", Owen King identifies a potential moral normative problem arising from many reasonable applications of image recognition software. Applied in particular to human persons and their visibly discernible behaviors, King argues that we should expect the moral problem of *presumption* to arise. If we think of the function of image recognition software as one of classification based on visual evidence and similarity relations, the presumption problem at its most general threatens insofar as *any* classificatory scheme fails to treat individuals as individual persons and thus fails to respect their moral status as such. Note at the outset that the problem of presumption is in no way unique to machine learning contexts. Indeed, King prefaces his discussion with a number of ordinary cases of human-on-human presumption, skillfully using concrete scenarios to guide intuitions about the moral problem of presumption. In the case of this general sort of presumption, consider the predilection we have with stereotyping, for one example, or our tendency towards confirmation bias, for another. More specific instances of presumption involve classificatory schemes grounded in illicit inferences to an individual's intentions. In the ordinary run of things we frequently must infer intention from behavior, including especially verbal behavior. Flubbing the inference, we react to the incorrectly attributed intention with (variously) resentment, dejection, confusion, humiliation, etc. Of course, image recognition software doesn't react, but it does classify and can be expected to be at least as fallible in the inferences drawn as we find ourselves to be. The problem comes in not discovering that presumption qua illicitly inferred intention has occurred and, as a result, the individual's autonomy is unduly restricted, albeit algorithmically.

King distinguishes between a *modular* approach to presumption and an *ingrained* approach. On the modular approach, cases of presumptive inference are (somehow) identified and excluded post classification, whereas the ingrained approach seeks to avoid the presumptuous classification in the first place. Rejecting the modular approach as the obviously question-begging alternative it is, King focuses his efforts in this contribution on how training data can be so restricted as to ensure "responsible judgments" – that is, non-presumptuous or at least minimally presumptuous classifications.

Responsible judgments are one problem, responsible *agents* quite another. As roboticists engineer increasingly sophisticated general applications systems, we confront the thorny problem of whether and how to assess their moral responsibility – viz., moral praiseworthiness or moral blameworthiness. In her engaging and well-argued "Robotic Responsibility", Anna Wilks explores a possible middle ground between two manifestly implausible, yet apparently exhaustive, views of robot moral responsibility. On one hand, we might view robots as either morally neutral, morally innocuous, or perhaps (at most) moral innocents, insofar in each case as they merely express the moral agency of their designers and users, being themselves at most simple tools. Surely, though, there is nothing simple about

the contributions a robot makes to its environment, operating as it does quite independently. On the other hand, we might view robots as fully morally responsible agents, which seems absurd both *prima facie* and especially after reflecting on the kantian conception of moral agency qua rational beings capable of authoring and motivating their own agency in light of recognizing the moral duty entailed by moral law. Note that the kantian account of moral agency is notoriously demanding. Less demanding accounts can be given, but in none of them does full robotic responsibility survive the fact of their having been through and through designed, engineered, programmed, and trained by moral agents who seek only to extend their own agency, and the kantian account is anyway Wilks' preferred starting point.

Wilks finds in Margaret Gilbert's work on joint commitment (Gilbert 2014) the grounds for a middle position between these two positions which holds that there is a sense of collective moral responsibility which is not strictly reducible to the moral responsibility of individuals acting in concert. Thus collective moral responsibility is neither a linear nor a diffuse – that is, some other functional – aggregate of individual assignments of responsibility in a group effort. Wide moral responsibility in this sense presupposes a collective moral agency to which individuals contribute their efforts. As Wilks notes, this necessarily stretches our ordinary conception of moral responsibility inasmuch as irreducibility entails a standalone notion of group moral responsibility. Individual contributions to collective moral agency need not, however, presume full *individual* moral responsibility for all of the members of the collective. At most, Wilks argues, some or even just one member must be fully morally responsible, while the rest require only a degree of intelligence and autonomy for their actions to count towards the collective agency and, thus, moral responsibility. As Wilks puts it, "[i]t is not necessary for the doctor to be also a nurse, and a social worker, and an extremely powerful computational machine. Why then should we require that the machine be a doctor or a social worker, or even a person? Each one contributes something as an individual, but the responsibility for the overall task is ascribed to the whole group – since the utlimate deliberation and actions taken involve the joint commitment of the collective."

Robot colleagues, as it were, cannot be viewed as genuine moral agents if our sense of moral agency is individual, but that does not exclude the necessity of viewing them as potentially important members of a moral community and contributing to communal moral responsibility in their various ways. Of course, much more needs to be said about the degree of intelligence and autonomy required to be so viewed as a member of the community and not merely a tool for its use, yet as Wilks concludes, we at least begin having the altogether necessary conversation of just how we should view the incorporation of sophisticated robotic systems in collective expressions of agency and, ultimately, in assessing group moral responsibility.

A narrow application of robotics which nevertheless carries broad social implications concern Jason Borenstein and Ronald Arkin in their "Robots, Ethics, and Intimacy: The Need for Scientific Research". Sketching the conceptual terrain as best as can probably be done given the nature of the application in question, the authors point to the dearth of answers to important questions regarding the

role of social robotics, particularly ones deliberately designed to emulate intimate relationships in such a way as to induce strong feelings in users of attachment and love. Although the prospect of roboticists inquiring seriously about the nature of intimate and loving relationships may strike one as peculiar, science fiction literature and film has long speculated that robots will eventually be so sophisticated as to be capable of perfectly imitating participants in intimate relationships. Still science fiction at this point, the prospect is made more pressing by the propensity humans have to adopt and form relationships – frequently very important social relationships – with non-human animals and, of greater relevance, inanimate objects. Construed as animate objects, robots are readily suited to exploiting this tendency, thereby impacting important aspects of human life, our capacities to value, care, form attachments, and even love.

By carefully articulating a number of important research questions, Borentsein and Arkin lay out an ambitious research agenda for roboticists, philosophers, psychologists, and sociologists to pursue in light of the progress on the engineering front of intimate robotics. The questions tend toward the consequentialist, asking after possible sources of utility and disutility in the application of robotic systems to socially intimate contexts. For example, what psychology (beliefs, desires, and intentions) can the use of intimate robots be expected to engender in the user? What of the user's well-being, psychological and otherwise, particularly in light of the possibility that intimate robots may tend to push out ordinary human relations? Consider in this regard the development of carebots to provide care and companionship for the elderly and infirm, which can only be expected to limit opportunities for human interaction. Perhaps more troubling, how will the prospect of forming intimate attachments with socially sophisticated robots impact our expectations, understanding, and perhaps even capacity for forming ordinary human relationships? The authors remind us that there is a dearth of research on these and many other questions besides, while also pointing out that robotics entrepreneurs will not be reticent to develop and exploit market niches where social robots will be welcomed, for good or ill.

Not all human interactions with robots entail (one way) intimacy or even continuous involvement. Indeed, most of us will only briefly interact with robots as they are deployed by developers, owners, and users on behalf of organizational – including government, corporate, and medical – interests. Frances Grodzinsky et al.'s "Applying a Social-Relational Model to Explore the Curious Case of hitchBOT" use the example of hitchBOT – the social-media 'hitchhiking' robot star whose summary destruction in Philadelphia seems once and for all to have settled the experimenter's question, "can robots trust humans?" – to argue that robot owners bear responsibility for robot-human interactions even when not present at those interactions.

As the authors explain, what interests them particularly about hitchBOT is that, unlike non-social robots without a shred of 'hooks' to encourage anthropomorphizing, including perhaps hospital delivery robots, vacuuming robots, or even the ubiquitous automatic teller machine, hitchBOT was specifically designed to induce friendly feelings and feelings of trust towards it. That is, if benign,

it was nevertheless designed to be deceptive, even though a fair part of that deception included a social media presence. Drawing on research on the moral dimensions of social robotics understood in terms of interactions and social roles, the authors specify the special obligations the designers of unaccompanied robots incur, particularly as the robot interfaces become increasingly sophisticated so as to converge on ordinary human interface – conversationally, say, or visually.

The question of whether and how to consider the moral status of robots need not, however, be solely grounded in terms of social-relational models. Migle Laukyte argues for an altogether different approach to these questions, one derived from considerations in environmental ethics, in his "Against Human Exceptionalism: Environmental Ethics and Machine Question". Specifically, Laukyte starts from the position in environmental ethics known as 'Deep Ecology', which denies any position of special moral privilege – such as being a person, say – in the complex ecological web. Thus Deep Ecology entails a kind of thorough-going ecological egalitarianism, although it is unclear whether the egalitarianism in question extends to geographic features like lakes, mountains, or fjords.

Laukyte makes an important point in noting that our 'environment' has long been, and is being with exponential rapidity, enriched with robots constituting more or less autonomous nodes in what can be viewed as an (albeit artificially constructed) ecological web. This stretches our ordinary understanding of 'ecology', and Laukyte's argument is, in part, to make plausible just such an extension so as to provide Deep Ecology purchase on the problem of the moral status of robots. Setting the stage, his focus is not so much on the obvious example of the autonomous robot reacting to and contributing however it may to the ecological web in question, but on a much wider notion of artificially intelligent agents, regardless of their engineering features or even whether such agents are physically instantiated in some specific robotic form or other.

Attempting to meet the obvious rejoinder, that this application of the central theses of Deep Ecology unacceptably distorts 'ecology' to include both natural and artificial agents, grounded in part perhaps by virtue of the fact that artificial agents are, at least, non-living, is front and center in the challenges Laukyte takes up. His argument here takes place on two fronts: First, the capacities of artificial agents – 'mindclones', as Laukyte dubs them – make them difficult to distinguish from natural agents given their success in mimicing behavioral reportoires; second, our ecology in any case has been subject to various substitutions and permutations by selective cultivation and breeding since the development of agriculture. Thus it would be arbitrary to exclude artificial agents from the ecological web, a point well worth considering regardless of any further claims on behalf of, or following from, Deep Ecology.

An issue Laukyte does not directly address is the construction of artificial – that is to say, virtual – environments as a whole, populated with artificial agents (non-player characters, or NPC's) and avatars of human agents. If the notion of an ecology can be stretched to include artificial agents, perhaps it is but one further step to admit an entire virtual environment as a wholly constructed ecosystem. Regardless, the normative features of those environments, particularly for individual human agents

represented by avatars in the virtual, is an area of considerable debate. In "The Ethics of Choice in Single-Player Video Games", Erica Neely takes up the puzzle of the moral status of actions in virtual environments, arguing that it is intelligible to speak of harms and benefits caused by the decisions of users (players) and designers alike because of the effects those actions have on them, inside or outside of the virtual environment in which the actions are taken.

Neely draws a distinction between the intravirtual (within game) effects a player's choices in game might have on him or her and the extravirtual effects of those same choices and the potential carryover into extravirtual choices. For one example long discussed in the popular hand-wringing over violent video games and first person shooters, consider that the brutalizing choices the video game player makes while immersed in a given virtual environment may lower social inhibitions to making harmful real-world choices. Neely's argument, however, is far more subtle than this sort of straight-line sorites.

The intravirtual choices a player might make in a virtual environment could encourage the player to entertain or make unethical choices in other contexts, depending in part on how designers of virtual environments encourage or discourage such choices, which in turn depends on the sort of rewards system the designers have built in to the virtual environment. Neely's point, however, is that virtual environment designers seek to make intravirtual choices as close in nature to their extravirtual counterparts as possible – to make them, in terms of the player's experience, 'real' choices with 'real' consequences. Virtual environments thus gain traction with players insofar as they exhibit realism in approximating the gravity of extravirtual choices for players. How well designers themselves grasp the moral import of the degree of such realism they manage to incorporate so as to engage their players raises the moral stakes of the creation and use of virtual environments. The stakes can be for moral ill, as Neely notes. Yet, importantly, she also argues that it can be for moral good, perhaps as players learn in the virtual environment more sophisticated methods of moral deliberation. The onus at least in part is on the sensitivity of designers to such issues, but it also rests with the game player and the lessons they draw from being immersed in the virtual environment.

1.7 Trust, Privacy, and Justice

Finally, the internet itself and the various social networks it contains are a long-standing source of normative puzzles, particularly as they are the perfect targets for big data collection, its harnassing by algorithmic analysis for purposes of pinpoint profiling, categorizing, and generalizing, and the subsequent exploitation of these analyses by private, corporate, and government interests. Just as we ourselves make use of the networks and services therein provided, those entities and individuals providing them make use of us, often in manipulative, exploitative ways which succeed in part by virtue of their relative invisibility from the network user's perspective. The wealth of scholarship in response has been nothing short

of a renaissance in the study and defense of human rights to rival that of the enlightenment, up to having numerous political rammifications. Front and center to these discussions are questions of identity, autonomy, privacy, trust, and justice.

In "Obfuscation and Good Enough Anonymity", Tony Doyle argues in favor of obfuscation – that is, the deliberate muddying of the informational waters, as it were, by the individual's use of misleading or ambiguous data. With characteristic clarity, Doyle draws a straight line from obfuscation to human well-being: Cleverly used, obfuscation can foil big data analytics in such a way as to preserve anonymity and thereby protect privacy as a way to defend, in turn, against manipulation and promote individual autonomy, where individual autonomy tends to promote individual well-being.

There are many caveats and exceptions to be drawn at each stage of Doyle's argument. The use of big data analytics need not be a zero-sum game. Consider their use as simply a matter of efficient and effective discrimination, and note that discrimination per se can be just or unjust, depending on the basis for discrimination. Moreover, since the data in big data analytics largely consists of the digital imprint ones online behavior in social and other networks makes, an argument can be made that the resulting discrimination neatly avoids the superficial bases – skin color, say, or height, or attractiveness – we otherwise tend in practice to use, wholly unjustly, to draw distinctions between people.

Yet given the fact of big data analytics and despite its many potential benefits, its service for altogether powerful and particular (though not necessarily malevolent, Doyle is quick to note) interests at the expense of broader social interests and specific personal interests tips the scale of the potential harms of obfuscation – of which there are many Doyle chronicles – in favor of the singular but overridingly important benefit of individual well-being. Or so Doyle argues. Of course, his is a practical as much as it is a philosophical argument. Doyle's prescription of obfuscation requires effort at evasion and something like subterfuge on the part of the network participant to secure even partial anonymity, at which proposal the tendency to throw ones hands up in surrender to the overwhelming force of big data analytics is understandable. After all, we've grown inured to a loss of privacy, just as we grow accustomed, horribly enough, to the potential and sometimes fact of exploitation made possible by our online presence. If 'going off the grid' is not feasible, as for most it is not, then perhaps Doyle has offered at least some line of defense.

Doyle's prescription to manage risk presupposes a broader account of the risk created by complicated online environments. Massimo Durante offers one such account in his "Trust and Security in the Digital Age: Algorithms, Standards, and Risks". Durante draws a crucial distinction between safety and security: Where safety is the immediate defense of life and well-being from threat, security protects ones life projects, including presumably their inception, fostering, and fruition. Frequently – and sometimes, perhaps, deliberately – confused, safety is a necessary condition on security, but not vice versa. The serf obtains safety, for example, but at the expense of security insofar as their life projects are their aristrocrat's, not their own.

Security is a uniquely critical feature of well-constructed online environments, since such environments have themselves become decidedly necessary to the projects of today's lives. Yet this puts the individual, unavoidably, in the position of delegating security to corporate entities and government agencies. The complex, altogether distributed nature of the online ecosystem presupposes, for the sake of security, risk-management at the levels of design, implementation, and application, with a particular and altogether necessary emphasis on automated risk-management and the development of trusted systems. Feeble libertarian fantasies aside, no individual has the capacity to ensure their own security in such an environment. Security, in short, presupposes trust, yet trust itself engenders risk. The design decisions made for purposes of risk-management in the development of trusted systems effectively codify and automate social values which, whether by intention or not, may give the appearance of transparency while nevertheless incorporating subtle discriminations and manipulations. The problem is all the more acute because many of the design, development, and implementation decisions are in turn opaque to democratic evaluation.

Durante points to a yet darker possibility: The common confusion of security for safety is ripe opportunity for exploitation by governments. After all, claiming threats to safety as justification for massive surveillance, data-harvesting, and data-analytics as per Doyle's argument dramatically impinges on security as Durante construes the distinction between safety and security. Safety, as a ploy, threatens security and with it the promise of online environments to play an integral role in life's projects. This is an ancient tension, to be sure, yet it is one made all the more pressing by the technology involved.

Ugo Pagallo closes the volume, appropriately enough, examining the legal-philosophical implications of hard legal cases emerging from the use of information technologies. Just as hard cases in ethics are useful to study because of the rift they expose between, say, utilitarian and deontological moral normative analyses, hard legal cases expose the gulf between tolerance-based and justice-based approaches in legal normative analysis – or, as Pagallo dubs them, lockean and platonic approaches, respectively. Complicating matters is the fact that some of the hard legal cases at the leading edge of law and politics regarding information technologies are genuinely novel and surprising, while many others simply continue traditions of posing long-standing legal puzzles and conflicts. That is, some are indeed new wine, while many are old wine in new bottles.

Nevertheless, hard legal cases of information technologies drive dispute among scholars in the first instance on whether a solution exists and, in the second instance, on just how the unique solution, or resolution, if it exists, is to proceed in weighing justice considerations against tolerance, and vice versa. Resolution, if attainable, reveals a legal paradox of sorts, since the cases depend on both the tolerant application of justice and the just use of tolerance, each setting limits on the other. Yet bouncing from tolerance to justice and back from justice to tolerance, otherwise separately at odds with one another in approach and outcome to legal hard cases, leaves their resolution an open question. Focusing on numerous examples from information ethics and jurisprudence, Pagallo argues for a nuanced methodological

analysis which shows one way by which a middle ground between justice and tolerance can be found, the one tempering the other in application to the hard cases.

1.8 Concluding Remarks

Harnessing computation from theory to engineering to application in its many permutations has clearly presented unique scholarly opportunities, all of them so interdisciplinary as to obliterate distinctions of discipline altogether. Are the philosophers roboticists, or are the roboticists philosophers? Are the mathematicians neuroscientists, or are the neuroscientists mathematicians? Are the legal theorists computer scientists, or are the computer scientists legal theorists?

In the end, and in light of the preceding discussions, the only reasonable response is a shrug: It just does not matter. As it has swept through every discipline, the computational turn has succeeded in wiping clean the deploringly artificial distinctions between those disciplines wrought by the balkanization of the resource-deprived modern university. Instead, threads of inquiry are woven throughout and link seemingly disparate research agendas, threads this introduction strives to highlight.

Their multi-disciplinary – better, *a-disciplinary* – investigations reveal the fruitfulness of erasing distinctions among and boundaries between formally established academic disciplines. This should come as no surprise: The computational turn itself is a-disciplinary, and no former discipline, whether scientific, artistic, or humanistic, has been left untouched. Rigorous reflection on the nature of these transformations, as we have seen, opens the door to inquiry into the nature of the world, what constitutes our knowledge of it, and our understanding of our place in it. That these investigations are only just beginning is signaled in part by the many contributions to this volume which close by describing open problems and inviting further research.

References

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, 609–626. Princeton: Princeton University Press.

Boolos, G., and J. Jeffrey. 1989. *Computability and Logic*, 3rd ed. Cambridge: Cambridge University Press.

Carnap, R. 1998. The value of laws: Explanation and prediction. In *Philosophy of Science: The Central Issues*, 2nd ed., ed. M. Curd and J. Cover, 678–684. New York: W.W. Norton.

Church, A. 1932. A set of postulates for the foundation of logic. *Annals of Mathematics (2nd Series)* 33(2): 346–366.

D'Agostino, M., and F. Luciano. 2009. The enduring scandal of deduction is propositional logic really uninformative? *Synthese* 167: 271–315.

Dennett, D. 1987. *The Intentional Stance*. Cambridge: The MIT Press.

Eliasmith, C., and O. Trujillo. 2014. The use and abuse of large-scale brain models. *Current Opinion in Neurobiology* 25: 1–6.

Fetzer, J. 2014. Information: Does it have to be true? *Minds and Machines* 14(2): 223–229.

Floridi, L. 2007. In defence of the veridical nature of semantic information. *European Journal of Analytic Philosophy* 3(1): 31–42.

Frege, G. 1980. On sense and reference. In *Translations from the Philosophical Writings of Gottlob Frege*, 3rd ed., ed. P. Geach and M. Black, 36–56. Oxford: Blackwell.

Gilbert, M. 2014. *Joint Commitment: How We Make the Social World.* Oxford: Oxford University Press.

Gödel, K. 1931. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für Mathematik und Physik* 33: 173–198.

Hempel, C. 1998a. Two basic types of scientific explanation. In *Philosophy of Science: The Central Issues*, 2nd ed., ed. M. Curd and J. Cover, 685–694. New York: W.W. Norton.

Hempel, C. 1998b. Two basic types of scientific explanation. In *Philosophy of Science: The Central Issues*, 2nd ed., ed. M. Curd and J. Cover, 808–825. New York: W.W. Norton.

Kramer, A.I., J.E. Guillory, and J.T. Hancook. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

Piccinini, G. 2017. Computation in physical systems. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. Stanford, Metaphysics Research Lab, Stanford University, summer 2017 edition.

Putnam, H. 1975. *Minds and Machines*, 362–386. Cambridge: Cambridge University Press.

Scarantino, A., and G. Piccinini. 2010. Information without truth. *Metaphilosophy* 41(3): 314–330.

Shannon, C.E. 1949. Communication theory of secrecy systems. *Bell System Technical Journal* 28(4): 656–715.

Turing, A. 1936. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society* 42: 230–265.

Part I
Computation and Information

Chapter 2
Computation in Physical Systems: A Normative Mapping Account

Paul Schweizer

Abstract The relationship between abstract formal procedures and the activities of actual physical systems has proved to be surprisingly subtle and controversial, and there are a number of competing accounts of when a physical system can be properly said to implement a mathematical formalism and hence perform a computation. I defend an account wherein computational descriptions of physical systems are high-level normative interpretations motivated by our pragmatic concerns. Furthermore, the criteria of utility and success vary according to our diverse purposes and pragmatic goals. Hence there is no independent or uniform fact to the matter, and I advance the 'anti-realist' conclusion that computational descriptions of physical systems are not founded upon deep ontological distinctions, but rather upon interest-relative human conventions. Hence physical computation is a 'conventional' rather than a 'natural' kind.

Keywords Computational Theory of Mind · Physical computation · Simple mapping account · Pancomputationalism · Computational stance

2.1 Introduction

What is computation? There are two basic ways to look at the issue: (1) in theory, as a type of *mathematical 'process'* – as something that belongs to a purely abstract and formal domain, like topology, set theory or real analysis; and (2) in practice, as the activity of certain *physical systems* – as what computers do, where a *computer* is a concrete device that exists in actual space and time. The connection between these two perspectives is generally conceived to lie in the *implementation* relation: a physical system or device performs a computation when

P. Schweizer (✉)
Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, Edinburgh, UK
e-mail: paul@inf.ed.ac.uk

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_2

it ‘implements’ or ‘realizes’ a particular abstract formalism. However, specifying the criteria under which the implementation relation properly obtains has proved surprisingly subtle and controversial, and there are a number of opposing views on the constraints that must be satisfied in order for a physical system to count as a ‘genuine’ implementation.

2.2 A Simple Mapping Account

A very straightforward and elegant account articulated by Putnam (1988) and others is based on a simple mapping between physical structure and abstract formalism. Accordingly, a physical system *P* performs a computation *C* just in case there is a *mapping* from the actual physical states of *P* to the abstract computational states of *C*, such that the transitions between physical states reflect the abstract state transitions as specified by the mapping. The minimalism, neutrality and generality of the Simple Mapping Account (henceforth SMA, adopting the terminology of Piccinini 2015a) make it a natural choice as the in-principle standard for physical implementation – it takes the Mathematical Theory of Computation (MTC) as its starting point and adds no substantial assumptions. And because it adds no further assumptions or restrictions, SMA is in an important sense maximally liberal – there will exist abstract mappings from a *huge* class of physical systems and processes to an equally huge class of computational formalisms.

And there is a clear sense in which this is a significant theoretical virtue. It is standard practice in computer science to apply computational descriptions to various physical systems *at will*, simply on the condition that the mapping yields an interesting or useful perspective. For example, simple physical devices such as parking ticket dispensers or traffic light controllers can be modelled in terms of Finite State Machines, without any reference to the original intentions of their designers nor to the actual details of their internal causal structure. In such cases, computational ascriptions constitute an *idealized* depiction, one that abstracts away from many actual features of the device to yield a simplified *formal model* of selected aspects of the device. This is quite analogous to applying mathematical formalisms such as differential equations to various physical systems to characterize aspects of their trajectories through state space. In both cases, the mapping from physical phenomena to mathematical formalism is highly reliant on both idealization and approximation, and deliberately neglects many aspects of the internal causal mechanisms.

In this type of endeavor, *which* aspects of the system are selected for abstract modelling is not fundamental to the system *per se*, but instead remains a question of human choice relative to our interests and goals. There are any number of different perspectives and levels for describing the very same system, and none of them is privileged. A traditional spring-driven analogue clock can be formally modelled at various *microphysical* levels – at a subatomic level in terms of quantum mechanical

processes and interactions, and at a higher microphysical level in terms of molecular thermodynamics. In the latter case, it could also be described in more abstract functional terms as a temperature detector, where the mean molecular kinetic energy of its metallic components tracks the ambient atmospheric temperature. And it can be described and modelled at various *macrophysical* levels as well, such as an intricate classical mechanism with states evolving in accord with continuous real valued equations. It could also be described in more idealized *conventional* terms, where certain selected continuous features are broken into discrete segments and given a chronological interpretation. And yet again, this relatively advanced design level stance could be ignored, and the object could be given a more rudimentary functional depiction, e.g. where its size and inertial properties make it useful as a doorstop.

For computation to remain an unfettered, and maximally adaptable mathematical tool, like set theory or topology, it is requisite that no fixed or preconceived limits be placed upon its potential range of physical interpretation. And indeed, SMA exemplifies this neutrality and universality with respect to the possible relations between abstract formal structure and 'concrete' physical phenomena. In this vein, Putnam (1988) gives a technical proof of the theorem that every open physical system implements every (inputless) Finite State Machine (FSM). He provides a generic depiction of a physical system as a bounded, continuous region of space-time, and the basic idea is that the region is held constant but sliced up in an as many different ways as one likes in order to define a sequence of disjunctive 'physical states' that can be mapped to any given run of a FSM.

And Searle famously promulgates the universality of SMA with the claim that virtually *any* physical system can be interpreted as implementing virtually *any* formal procedure. For example, Searle (1990) asserts that the molecules in his wall could be interpreted as running the WordStar program. The claim is simply put forward with no further defense, but Copeland (1996) provides a proof of what he calls 'Searle's Theorem', which he observes is essentially a notational variant of Newman's (1928) objection to Russell (although Copeland then goes on to reject SMA).

This broad-minded position on physical computation arises as the natural inverse of the standard and uncontroversial view that abstract formal procedures, as such, are *multiply realizable*. It's clearly possible to implement the *very same* computational formalism using vastly different arrangements of mass/energy. Following notation and terminology introduced in Schweizer (2012), let us call this top-down feature 'downward multiple realizability', wherein, for any given formal procedure, this *same* abstract formalism can be implemented via an arbitrarily large number of *distinct types* of physical systems. And let us denote this type of downward multiple realizability as '↓MR'. The basic perspective advocated by Putnam and Searle then goes in the reverse direction. Let us call the bottom-up view that any given *physical system* can be interpreted as implementing an indeterminately large number of different *computational formalisms* 'upward MR' and denote it as '↑MR'. The basic import of ↑MR is the non-uniqueness of computational

ascriptions to particular configurations of mass/energy. In the extreme versions of ↑MR propounded by Putnam and Searle, it is not simply a case of non-uniqueness, but rather there are apparently no significant constraints at all – it is held to be possible to interpret virtually any open physical system as realizing virtually every computational procedure. Let us call this more extreme version '*universal* upward MR' and denote it as '↑MR*'. ↑MR* is noteworthy in that it provides the theoretical limit case in terms of abstraction away from physical specifics or limitations, and in this sense is comparable to the idea that, e.g., any physical object can be an element in a limitless number of distinct sets.

2.3 The Computational Stance

Many philosophers have found the degree of liberality induced by SMA objectionable. Historically, these objections stem from the conflict between critics *versus* proponents of the Computational Theory of Mind (CTM). Critics of CTM have used SMA to argue that a computational approach to the mind is empirically vacuous. These 'trivialization' arguments hold that, *a la* ↑MR*, a mapping will obtain between virtually any physical system and virtually any formalism, which in turn is construed as fatally undermining CTM, since whatever computational procedures are held to account for our cognitive attributes will also be realized by a myriad of other 'deviant' physical systems, such as buckets of water and possibly even stones. Hence by CTM it would seem to follow that such obviously insentient systems have the same cognitive attributes that we do, which is then taken as a *reductio ad absurdum* disproof of CTM.

In response to ↑MR* and the associated trivialization claims, a host of authors, including Fodor (1981), Maudlin (1989), Chrisley (1994), Chalmers (1996), Copeland (1996), Shagrir (2001), Block (2002), Sprevak (2010), Milkowski (2013), Rescorla (2014), Piccinini (2015b) advocate additional constraints on the implementation relation, so that it is no longer a 'simple' or theoretically neutral mapping. In effect, these restrictions serve to preclude a vast number of physical systems from the domain of the mapping function, in an attempt to separate 'true' or 'genuine' implementations from the many presumably 'false' cases countenanced by SMA. These constraints include: causal, counterfactual, semantic, and mechanistic/functional criteria.

However, I advance quite a different type of response to the situation. First, instead of attempting to 'save' CTM by constraining the account of physical implementation, I hold that SMA does not actually constitute a threat to scientifically plausible versions of CTM. No one thinks that SMA 'threatens' electrical engineering or our ability to design and utilize sophisticated computational artifacts, and in my view, the particular version of CTM that *is* undermined by SMA is not one that should be accepted in any case. Second, I argue that none of the proposed constraints provides a truly general and satisfactory 'realist' account of

physical implementation – indeed, none succeeds at providing a globally applicable necessary condition. So I advocate retaining a very liberal SMA view of physical implementation, that derives from the basic insights of Turing, Kripke, Putnam and Searle, while rejecting the standard anti-CTM conclusion of the trivialization arguments.

In line with SMA and ↑MR, I argue that computation is not an 'intrinsic' property of physical systems, in the sense that (a) it is founded on an observer-dependent act of ascription, upon an entirely *conventional* correlation between physical structure and abstract formalism. Furthermore, (b) this conventional mapping is essentially prescriptive in nature, and hence projects an outside *normative* standard onto the activities of a purely physical device. In this manner we adopt what could be termed a 'Computational Stance' towards physical systems. This approach is in some ways comparable to Dennett's (1981) Intentional Stance, wherein intentional states such as beliefs and desires are not posited as objectively real phenomena. Instead, they are treated as mere 'calculational devices' or '*abstracta*' in Reichenbach's sense (like point masses and perfectly frictionless surfaces in classical mechanics), used to predict observable events but without any additional ontological commitments.

Analogously, I would construe abstract computational states on a similar footing. In the case of our purpose-built artifacts, these abstract states are *idealized* formal notions that we employ to describe such devices from a higher design-level perspective. Classic digital computation is rule-governed syntax manipulation, and as such is no more intrinsic to physical configurations than is syntax itself. Furthermore, discrete states are themselves idealizations, since the physical processes that we interpret as performing digital computations are continuous (in the standard non-quantum case). Thus discrete states do not literally correspond to the underlying causal substrate. We must *abstract away* from the continuity of actual physical processes and impose a scheme of conventional demarcations to attain values that we can then *interpret* as discrete. Hence this elemental building block of digital procedures must be projected onto the natural order from the very beginning (as Turing observed in 1950), and in this respect is a convenient fiction rather than a literal depiction.

Dennett holds that there is no internal matter of fact distinguishing systems that 'really' possess intentional states from those which do not – the strategy only requires us to view the system *as if* it possessed such states. Hence there is nothing in principle to stop one from depicting a stone as an intentional system if one so chooses. In a similar vein, I would argue that there is no deep or metaphysically grounded fact regarding whether or not a physical system 'really' implements a given computational formalism. In the case of artifacts such as my desk top computer, I can gain a huge increase in the ability to predict its future states if I adopt a computational stance as opposed to viewing it as a brute physical mechanism. And this is because it has been designed and constructed for exactly this purpose, just as an electric toaster has been designed and constructed to perform a particular function. In contrast, a stone has not been so designed, and the pragmatic value of viewing it in computational terms will be rather limited.

2.4 Critique of the Causal Account

I will now critically address some of the proposed constraints on SMA, with the aim of showing that none provides a principled necessary condition for physical computation. The causal account supplies one of the most natural and intuitively compelling constraints on the implementation relation. Chalmers (1996), for example, contends that it is a necessary condition that the pattern of abstract state transitions must be the image under the mapping of an appropriate transition of physical states of the machine, where the relation between succeeding physical states in this sequence is governed by *proper causal regularities*. Furthermore, these regularities are supposed to 'mirror' the structure of the abstract formalism. The imposition of such a constraint will screen off a vast number of Putnam's sequences of physical states, with the aim of reducing the domain of the mapping function to a tiny subset of purportedly 'legitimate' cases of implementation.

However, I argue that the causal constraint is too stringent in general and rules out a significant number of cases which should not be excluded. And although it's a more specialized and sophisticated approach, the mechanistic/functional account shares some key features with the causal, so many of the following criticisms carry over to this account as well. A basic problem with causal and mechanistic approaches is that they place emphasis on the wrong level of conceptual analysis. Rather than addressing the question of *whether or not* a given configuration of mass/energy implements a given computational formalism, causal considerations instead address the lower level and divergent practical question of *how*, in certain circumstances and over limited spans of time, this implementational sequence is mechanically generated.

The inessential status of causal structure can be elucidated with the observation that the key factor in judging *that* a given configuration of mass/energy implements a particular computational formalism is *simply because*, according to our abstract blueprint, the *correct* series of physical state transitions actually occurs. As an exemplary case of where appeal to causal regularities is completely irrelevant to determining whether or not a given sequence of states counts as an implementation, consider Turing's original (1936) heuristics, where the paradigm of actualized computation is a *human computor*, meticulously following a program of instructions and executing computations by hand with pencil and paper. In this seminal and classic example of concrete realization, the transitions from one state to the next are not governed by causal regularities in any straightforward mechanical sense. When I take a table of instructions specifying a particular abstract TM and perform a computation on some input by sketching the configuration of the tape and read/write head at each step in the sequence, the transitions sketched on the paper are *not themselves* causally connected: as in the virtual machine states in standard computers, one sketch in the sequence in no way causes the next to occur.

In terms of the 'causal' factors underpinning their occurrence, it is primarily through my understanding of the instructions and intentional choice to execute the procedure that the next stage in the sequence appears. But my complex behaviour as a human agent deliberately following instructions is not something that we currently have any hope of being able to recast in terms of causal regularities at the purely physical level of description. Furthermore, whatever causal factors at this level *do* ultimately underwrite my ability to execute the procedure, they will be exceedingly convoluted and indirect, and there is no reason to believe that they will 'mirror' (or even remotely resemble) the structure of the formalism. In cases of *intentionally mediated* causation, we accept the sequence of configurations on the paper as an implementation of the program, not because we have the faintest idea of the underlying causal story, but rather because the sequence itself is *correct* and can be seen to follow the procedural rules. In other words, the projected mapping, *a la* SMA, has been preserved.

To continue the example, consider the following 3 state Turing machine ***M*** given by the four quadruples:

$$q_1 1 R q_1 \quad q_1 0 1 q_2 \quad q_2 1 L q_2 \quad q_2 0 R q_3$$

The first element in each quadruple (e.g. q_1 in the first case) is the current state, the second element is the currently scanned symbol (either 1 or 0) the third element is the overt action (*move* R or L one square, or *print* a 1 or a 0), and the last element is the covert 'act' of entering the next state. Now suppose I'm confronted with an initial tape configuration

$$01100\ldots \quad \text{(all other squares to the right are blank)}\ .$$

Armed with the foregoing explication of the quadruple notation, along with a few basic operational conventions (as described in Boolos and Jeffrey 1989), I can act as a perfectly good human computor and manifest a physical implementation of the respective Turing machine computation. With pencil and paper I can perform the sequence of 6 transitions determined by the input configuration and then halt. Indeed, I've now keyed into the digital file generating this document the very same sequence that I sketched in my notebook, and have thus produced an alternate physical realization of the same computation. The machine starts in its lowest numbered state reading the leftmost non-blank square (where the contents of a square are indicated by the corresponding digit in the tape string). An underline indicates the currently scanned square, and the number below this indicates the current state. The machine halts when it enters a state for which there is no instruction.

$q_1 1 R q_1$; $q_1 0 1 q_2$; $q_2 1 L q_2$; $q_2 0 R q_3$ on input 01100…

Start 0$\underline{1}$100…
1
01$\underline{1}$00…
1

01$\underline{0}$0…
1
011$\underline{1}$0…
2
01$\underline{1}$10…
2
0$\underline{1}$110…
2
$\underline{0}$1110…
2
0$\underline{1}$110… Halt
3

It's important to note that the foregoing sequence of configurations is not just a *linguistic description of* a possible physical implementation. Instead, the actual syntactic tokens are *themselves* concrete realizations extended in physical space-time. Manipulating syntactic tokens on a piece of paper *is* a transformation of the physical environment that itself constitutes a realization of the abstract formalism. And the same is true of the sequence of symbols generated above – it's a physical implementation of the abstract TM computation generated by Microsoft Word.

But what is the *causal structure* underling the Microsoft implementation? It doesn't really matter. The entries in this sequence bear no decipherable *causal* relations to each other – they're simply generated by what is stored in the digital file that is stored in the computer connected to the monitor. The actual computation in space-time appearing as I type is a sequence of illuminated patterns projected onto the screen, not supported by any causal regularities that 'mirror' the structure of the Turing machine program. It's surely true that every event must have a cause, but my point is that *surface inspection alone* reveals that this sequence is a proper realization of the specified TM program on input 01100 . . . To arrive at the judgement, we do not need to know *anything* about the causal mechanisms whereby this sequence was produced.

And what is the *semantic interpretation* of the Microsoft implementation? Again, it doesn't really matter. The computation itself is comprised of rule governed syntactic transformations. How these transformations are then semantically construed is superfluous to the execution of the program. If we choose, we *can* interpret $\boldsymbol{M}$'s activity as computing the function $f(x) = x + 1$ on positive integers expressed in monadic notation (and which halts on the same square at which it starts), so that the foregoing sequence of configurations is a computation of $f(2) = 2 + 1 = 3$. However, this is clearly not essential to the formal procedure itself.

And what *would have happened if* a different input string had been attempted? Again, it doesn't really matter. What matters is that, in accord with the formal procedure, the foregoing sequence is *correct* – it satisfies the essential normative specification as a series of rule governed transformations on the input specified.

2.5 Implementation as Proof in First-Order Logic

Each quadruple in the TM program can be seen as a conditional instruction, so that, e.g., the first quadruple is the conditional: *if* in state q_1 reading a 1, *then* print a 0 and enter state q_1. Hence it is the *logical* form of the if-then statement that captures the significance of the TM instruction, and this is all that must be satisfied by an implementation. Again, this is a quintessentially *normative* constraint, and it's a basic fact of logic that the truth-functional character of the material conditional does not imply *any causal connection* between antecedent and consequent.

This same fundamental point is made even more graphic by noting that Turing machine computations can be formalized in first-order logic with identity (FOL=). Each quadruple instruction can be rendered as a *universally quantified conditional* indicating the result of executing the instruction. In providing the details of the formalization, our object language $\boldsymbol{L}$ for FOL = will contain the symbols $\mathbf{o}$ and $'$ as distinguished vocabulary items, where $\mathbf{o}$ is a singular constant that, under the intended interpretation $\mathcal{I}$, denotes the number 0, and where $'$ is a 1-place function symbol which under $\mathcal{I}$ denotes the successor function. With these resources we can construct canonical numerals intended to denote numbers in the obvious fashion, e.g., where $\mathbf{o}'$ is the numeral for the number 1, $\mathbf{o}''$ the numeral for 2, etc.

In order to formalize the very simple machine $\boldsymbol{M}$ depicted above, we can make do with the assumption that the operant squares are unbounded only to the right. Furthermore, a blank square is construed as containing the symbol '0', and only finitely many squares are ever non-blank (i.e. contain the symbol '1'). To begin the formalization, let all the operant squares of the tape be labelled by a natural number, with the leftmost such square labelled with 0, the next with 1, etc. (the labelling number is distinct from the symbol occurring in the square). We adopt the convention that the positive integer input is expressed in monadic notation, with the leftmost '1' occurring in square number 1. At the start of the computation, all non-input squares of the tape are blank, and the machine starts in state 1 reading square 1.

Let t be the 'time' variable ranging over steps in the computation. We need two final FOL vocabulary items: for each state q_i of a given machine, pick a 2-place predicate $\mathbf{Q}_i$. For each symbol S_j the machine can read/write, pick a 2-place predicate $\mathbf{S}_j$ (in this case there are only two). The domian $\mathcal{D}$ of the intended interpretation $\mathcal{I}$ is the set of natural numbers, and $t\mathbf{Q}_i x$ is true in $\mathcal{I}$ iff at time t $\boldsymbol{M}$ is in state q_i scanning square number x, and $t\mathbf{S}_j x$ is true in $\mathcal{I}$ iff at time t the symbol S_j is in square number x. With these details in place we can now proceed to formalize $\boldsymbol{M}$'s program of instructions.

The first quadruple $q_1 1 R q_1$ is rendered as the 'axiom' **A1**

$$\forall t \forall x \forall y \left[(t \mathrm{Q}_1 x \wedge t \mathrm{S}_1 x) \rightarrow \left(t' \mathrm{Q}_1 x' \wedge \left(t \mathrm{S}_0 y \rightarrow t' \mathrm{S}_0 y \wedge t \mathrm{S}_1 y \rightarrow t' \mathrm{S}_1 y \right) \right) \right]$$

Under the intended interpretation this axiom 'says that' *if* machine ***M*** is in state q_1 at time t scanning square number x on which the symbol S_1 (= 1) occurs, *then* at time $t + 1$ ***M*** is in state q_1 scanning square $x + 1$, and in all squares the same symbol appears at time $t + 1$as at time t.

Various authors (including Chalmers 1996 and Copeland 1996) have objected to Putnam's proof because it relies on material conditionals, and it is claimed that more powerful *counterfactual* machinery is required to account for possibilities other than the input actually given. However, it is significant to note that the above universally quantified conditional ranges over *all times* and *all squares* in *any* computation, and hence exhaustively covers all relevant possibilities.

$q_1 0 1 q_2$ is rendered as **A2**

$$\begin{aligned} &\forall t \forall x \forall y \left[(t \mathrm{Q}_1 x \wedge t \mathrm{S}_0 x) \right. \\ &\quad \left. \rightarrow \left(t' \mathrm{Q}_2 x \wedge t' \mathrm{S}_1 x \wedge \left(\mathrm{y} \neq \mathrm{x} \rightarrow \left(t \mathrm{S}_0 y \rightarrow t' \mathrm{S}_0 y \wedge t \mathrm{S}_1 y \rightarrow t' \mathrm{S}_1 y \right) \right) \right) \right] \end{aligned}$$

$q_2 1 L q_2$ yields **A3**

$$\forall t \forall x \forall y \left[\left(t \mathbf{Q}_2 x' \wedge t \mathbf{S}_1 x' \right) \rightarrow \left(t' \mathbf{Q}_2 x \wedge \left(t \mathbf{S}_0 y \rightarrow t' \mathbf{S}_0 y \wedge t \mathbf{S}_1 y \rightarrow t' \mathbf{S}_1 y \right) \right) \right]$$

$q_2 0 R q_3$ yields **A4**

$$\forall t \forall x \forall y \left[(t \mathrm{Q}_2 x \wedge t \mathrm{S}_0 x) \rightarrow \left(t' \mathrm{Q}_3 x' \wedge \left(t \mathrm{S}_0 y \rightarrow t' \mathrm{S}_0 y \wedge t \mathrm{S}_1 y \rightarrow t' \mathrm{S}_1 y \right) \right) \right]$$

The set {**A1**,**A2**,**A3**,**A4**} **formalizes** ***M***'s program.

Next two arithmetical axioms are needed to govern the behavior of $'$ and **<**. The first axiom says that each integer is the successor of exactly one integer: $\mathbf{A}'$

$$\forall z \exists x \left(z = x' \right) \wedge \forall z \forall x \forall y \ \left(\left(z = x' \wedge z = y' \right) \rightarrow x = y \right).$$

The axiom governing **<** states that: **A<**

$$\begin{aligned} &\forall x \forall y \forall z \ (x < y \wedge y < z \rightarrow x < z) \wedge \forall x \forall y \left(x' = y \rightarrow x < y \right) \\ &\quad \wedge \forall x \forall y \, (x < y \rightarrow x \neq y) \ \text{(needed for the entailment relation below)} \end{aligned}$$

Finally, for the initial configuration with '01100' as starting input

$$(t = 0 \text{ in state } \mathrm{q}_1 \text{ reading square } 1) : \mathbf{A0}$$

$$\mathbf{oQ}_1\mathbf{o}' \wedge \mathbf{oS}_1\mathbf{o}' \wedge \mathbf{oS}_1\mathbf{o}'' \wedge \forall y \ \big((y{\neq}\mathbf{o}' \wedge y{\neq}\mathbf{o}'') \rightarrow \mathbf{oS}_0 y\big)$$

Let $\Delta = \{\mathbf{A1},\mathbf{A2},\mathbf{A3},\mathbf{A4}, \mathbf{A}',\mathbf{A<},\mathbf{A0}\}$

Now Δ completely formalizes the 'actions' of machine ***M*** on input '01100', and each step n in the previously sketched sequence of configurations, constituting the computation on input '01100', is s*yntactically encoded* by a sentence $\mathbf{T}_n$ in FOL=. Furthermore, the sentence $\mathbf{T}_n$ is ***logically entailed*** by Δ.

For $t = 1$ the sentence $\mathbf{T}_1$:

$$\mathbf{o}'\mathbf{Q}_1\mathbf{o}'' \wedge \mathbf{o}'\mathbf{S}_1\mathbf{o}' \wedge \mathbf{o}'\mathbf{S}_1\mathbf{o}'' \wedge \forall y \ \big((y{\neq}\mathbf{o}' \wedge y{\neq}\mathbf{o}'') \rightarrow \mathbf{o}'\mathbf{S}_0 y\big)$$

For $t = 2$ the sentence $\mathbf{T}_2$:

$$\mathbf{o}''\mathbf{Q}_1\mathbf{o}''' \wedge \mathbf{o}''\mathbf{S}_1\mathbf{o}' \wedge \mathbf{o}''\mathbf{S}_1\mathbf{o}'' \wedge \mathbf{o}''\mathbf{S}_0\mathbf{o}''' \wedge \forall y \ \big((y{\neq}\mathbf{o}' \wedge y{\neq}\mathbf{o}'' \wedge y{\neq}\mathbf{o}''') \rightarrow \mathbf{o}''\mathbf{S}_0 y\big)$$

For $t = 3$ the sentence $\mathbf{T}_3$:

$$\mathbf{o}'''\mathbf{Q}_2\mathbf{o}''' \wedge \mathbf{o}'''\mathbf{S}_1\mathbf{o}' \wedge \mathbf{o}'''\mathbf{S}_1\mathbf{o}'' \wedge \mathbf{o}'''\mathbf{S}_1\mathbf{o}''' \wedge \forall y\big((y{\neq}\mathbf{o}' \wedge y{\neq}\mathbf{o}'' \wedge y{\neq}\mathbf{o}''') \rightarrow \mathbf{o}'''\mathbf{S}_0 y\big)$$

$$\vdots$$

For $t = 7$ the sentence $\mathbf{T}_7$:

$$\mathbf{o}''''''\mathbf{Q}_3\mathbf{o}' \wedge \mathbf{o}''''''\mathbf{S}_1\mathbf{o}' \wedge \mathbf{o}''''''\mathbf{S}_1\mathbf{o}'' \wedge \mathbf{o}''''''\mathbf{S}_1\mathbf{o}'''$$
$$\wedge \forall y \ \big((y{\neq}\mathbf{o}' \wedge y{\neq}\mathbf{o}'' \wedge y{\neq}\mathbf{o}''') \rightarrow \mathbf{o}''''''\mathbf{S}_0 y\big)$$

M has no instructions for q_3 and hence will halt if it enters this state. So the 'canonical' Halting Sentence **H** for this machine is

$$\exists t \exists x \ \ (t\mathbf{Q}_3 x \wedge t\mathbf{S}_0 x) \vee \exists t \exists x \ \ (t\mathbf{Q}_3 x \wedge t\mathbf{S}_1 x)$$

and it's provable (by mathematical induction) that $\Delta \vDash \mathbf{H}$, since $\Delta \vDash$ T8 and T8 $\vDash$ **H**.

Logical entailment is an abstract mathematical relation, but a *particular proof* is a concrete syntactic phenomenon extended in physical space-time. In this manner, the foregoing Turing machine computation is equivalent to a proof in FOL=, and any such proof carried out with pencil and paper, following the rules of your favorite first-order deductive system, counts as a *physical implementation* of the computation.

It seems a very strange and implausible view to maintain that the property of being a proof in first-order logic is constrained by underlying causal regularities or mechanistic features. Indeed, when I mark student exams in my *Introduction to Logic* course, considerations of underlying causal regularities and biological mechanisms play no role whatever in determining whether some sequence of formulas is or is not a proof. The only thing that matters is whether or not the rules have been correctly followed, and this is a purely normative consideration. And since a proof of the relevant sort counts as an implementation of a Turing machine computation, it follows that causal regularities likewise have no bearing on the status of such implementations. Indeed, part of the reason that underlying causal considerations are the wrong level of analysis is that there is no sense in which error or malfunction can occur when viewed from this basic physical perspective. This thread will be resumed in Sect. 2.7.

The foregoing *counterexamples* show that causal and mechanistic factors do not impose a necessary condition on physical implementation. Instead, the *only* necessary condition is that the intended mapping, *a la* SMA, is preserved. In particular, we don't need to take into account the mechanics of *how* this success has been achieved in order to judge *that* it has it has occurred. And indeed, this is directly comparable to other abstract, rule governed activities such as chess. A game of chess is constituted by a sequence of moves on a geometrically defined board. Like computations, chess games are substrate neutral and can be realized in a virtually limitless variety of physical media. Furthermore, in ascertaining whether a given sequence is a legitimate game, all we need to know is whether or not each move is in accordance with the abstract structural rules of chess. The question of *how* these moves were physically accomplished is entirely irrelevant. Was the white bishop picked up and moved with the right hand or the left? Held between thumb and forefinger or thumb and index finger. Or perhaps moved by the power of psycho kinesis? Obviously the answer makes no difference.

2.6 Counterfactual Constraints

The *counterfactual* requirement is aimed at another apparently 'slack' feature incorporated by Putnam and the SMA, *viz*. the mapping from formalism to physical system is defined for only a single run, and says nothing about what *would* have happened *if* a different input had been given. And it is objected that this is too weak to satisfy the more rigorous operational notion of being a 'genuine' realization. However, in response to this quite natural proposal, it is worth noting that for a

physical system to realize a rich computational formalism with proper input and output capacities, such as an abstract TM, this will always be a matter of *mere approximation*. For example, any given physical device will have a finite upper bound on the size of input strings it is able to process, its storage capacities will likewise be severely limited, and so will its actual running time. In principle there are computations that formal TMs can perform which, even given the fastest and most powerful physical devices we could imagine, would take longer than the lifespan of our galaxy to execute. Hence even the fastest and most powerful physical devices we could envision will still fail to support *all* the salient counterfactuals.

So it will never be possible to construct a complete physical realization of an abstract TM – the extent to which a concrete device can execute the full counterfactual range of state transitions of which the abstract machine is capable will always be a matter of *degree*. For example, consider the exceedingly simple machine ***M*** given above. It's a straightforward matter to exhibit *particular* computations on small inputs. But there is no finite upper bound on the size of input strings that this abstract machine can handle. The set of four quadruples yields a mathematically well defined and effective procedure for adding one to a monadic input string which contains in excess of, say, $10^{10000000000000000000000000000}$ 1's. It's not physically possible for *any* artefact that we could build to carry out computations for such astronomically large inputs. Hence *no* physical implementation of this simple three state TM can deal with the full range of *possible* inputs.

So, in general, the class of counterfactual cases on alternative inputs with which a physical realization can cope is by necessity limited – not all counterfactual cases will be supported by *any* physical device implementing any TM. And this renders the appeal to counterfactuals inescapably *ad hoc.* The restrictive strategy demands that the mapping be able to support counterfactual sequences of transitions on inputs not actually given – but precisely *how many* inputs not actually given? One, two, twenty trillion? For any implementation, there will be a finite upper bound on the size of input string it can process, and beyond that size there will be *infinitely many* potential inputs for which it will not be able to perform the salient computation.

This indicates that there is no clear or principled cut off point demarking 'genuine' implementations from 'false' ones in terms of counterfactuals. As another, more common place, illustration of the *ad hoc* nature of the appeal to counterfactuals, consider a standard pocket calculator that can intake numbers up to, say, 6 digits in decimal notion. Is this a 'false' realization of the corresponding algorithm for addition, since it can't calculate $10^6 + 10^6$? It's an *approximate* instantiation which is nonetheless exceedingly useful for everyday sums. It will always be a matter of degree how many counterfactuals can be supported, where a single run on one input is the minimal case. Where in principle can the line be drawn after that? It's a matter of our purposes and goals as interpreters and epistemic agents, and is not an objective question about the 'true' nature of the physical device as an implementation. In some cases we might only be interested in the answer for a single input, a single run.

In addition, Bishop (2009) has importantly extended the SMA strategy to show that any predetermined finite set of counterfactuals *can* be accommodated on this

approach. From this I would conclude that the underlying and more general constraint of concern to those who would delimit the range of physical implementation is neither causal nor counterfactual. Instead, the point to emphasize is that in ↑MR* exercises of this sort, the mapping is entirely *ex post facto*. The abstract procedural 'trajectory' is already known and used as the basis for interpreting various state transitions in the open system and hence characterizing it as an implementation. Hence using this *ex post facto* tactic, even finite sets of counterfactuals can be included. And as emphasized above, our actual computational artefacts are themselves only capable of handling finite sets of counterfactuals.

For a physical device to successfully '*perform* a computation' is distinct from '*fully implementing* a computational formalism'. Performing a computation is an occurrent series of events, an actual sequence of physical state transitions yielding an output value in accord with the normative requirements of the mapping. And this can be satisfied in the case of computing the value of a single output on a given input. In contrast, fully implementing a computational formalism is a much more stringent and hypothetical notion, requiring appeal to counterfactuals, and as above, this will only ever obtain as a matter of degree. In light of this distinction, it is clearly possible for a physical device to successfully perform a computation *without* instantiating a complete computational formalism, which distinction in turn fatally undermines the theoretical force of counterfactuals in attempting to determine whether a physical process has 'really' performed a computation.

2.7 Computational Ascriptions Are Normative

As mentioned above, part of the reason why underlying causal considerations are the wrong level of analysis is that there is no sense in which error or malfunction can occur at this basic physical plane. Physical systems, as such, are governed by *natural laws*, while formal systems are intrinsically *rule governed*. In the case of our computational artefacts, a system governed by natural laws must be deliberately engineered so that it can be interpreted as evolving in accordance with a chosen rule governed formal system. 'Obedience' to natural law is an essentially *descriptive* matter and there is no sense in which mistakes or error can be involved – such laws cannot be broken, and the time evolution of material systems is wholly determined (in the classical case at least) by the regularities in question. On the other hand, 'obedience' to formal rules is an essentially *normativ*e matter, and there is a vital sense in which error and malfunction can occur.

This normativity has nothing to do with ethical or religious considerations, but simply with *conventionally imposed* norms. Suppose we are playing a game of chess. It's my move and it's clear that I'm about to be checkmated by your queen. So I pick up your queen and throw it out the window. You object with the exclamation 'You can't do that!' And I reply, 'What do you mean – I just did'. In this case the physical processes in question are in perfect accord with natural law, but have discontinued implementing the norms of chess. Similarly, if my desk

top machine is dosed with petrol and set on fire while still in operation, the time evolution of the hardware will remain in perfect descriptive accord with natural law. However, it will very soon fail to comply with the normative requirements of implementing Microsoft Word, and serious computational malfunctions will ensue. Being an implementation of Microsoft Word is a normative and *provisional* interpretation of the hardware system, which can be withdrawn when something goes 'wrong' or when the system is disrupted by non-design intended forces – being an implementation of Microsoft Word is not intrinsic to the physical structure itself. It is only at a *non-intrinsic* prescriptive level of description that 'breakdowns' can occur, and we characterize these phenomena as malfunctions only because our extrinsic ascription has been violated (as in Kripke 1982).

Accordingly, I would argue that the status of computation is very different than the status of abstract mathematical theories in physics. In physics we are attempting to give a fundamental characterization of 'reality', and in principle at least all existent phenomena supervene upon this fundamental level. There is no substrate neutrality in this case, and instead we are attempting to arrive at a theoretical description of the fixed and given natural order. So the mapping from abstract formalism to physical values is not purely conventional as with SMA – e.g. the variables are mapped to basic physical magnitudes and not just anything we please. And in the mathematical descriptions of basic physical theory there is *no normativity involved*. If the predictions of a particular theory, say Newtonian mechanics, turn out to be incorrect in certain cases, we do not say that physical reality has therefore 'malfunctioned'. Instead we say that Newtonian mechanics is at fault and our mathematical description *itself* is incorrect.

Imagine that we take a device intended to compute some given arithmetical function. There is always a non-zero probability of error for any algorithm implemented in the physical world – files become 'corrupted', overheating induces processing 'faults', 'errors' are propagated. Since error is always possible it follows that there is no independent fact of the matter regarding which function or algorithm is 'really' being computed. Suppose we say that the device is computing addition. We confirm this by testing its behaviour on 50 thousand inputs and it gives the correct outputs. But unknown to us the device possesses a mechanical fault, and when we keep going it gives some 'wrong' answers for larger inputs. So which function is it *really* computing – addition with errors, or the actual function in extension that corresponds to its physical behaviour? I would say there is no objective fact to the matter. In the arithmetical case there's an extra level of attributed *abstract* computational 'behaviour' that is always *underdetermined* by its actual performance, and which does *not* supervene upon underlying physical microstructure.

According to Piccinini (2015b), one of the prime advantages of the mechanistic approach is that it can account for cases of miscomputation. In this regard it diverges from a purely causal story by invoking normative/functional considerations. However, I would respond that these normative standards are not objective features of physical systems *per se*, but rather are purely conventional human interpretations, on the same par with computational ascriptions themselves. In the case of artifacts, the

mechanistic account must invoke the intentions of the human designers in order to characterize error and malfunction. But this does not successfully address Kripke's philosophical critique, since the purpose and normativity are still entirely in the eye of the human beholder.

In the case of biological systems, including brains, the mechanistic account shifts the burden of the intentional homunculus onto the 'purposiveness' of biological 'design'. According to this type of neo-Darwinian strategy, something has a particular biological function if this function was selected in the course of the organism's evolutionary history. In the present discussion there is not sufficient space to offer a sustained critique of this move. However, in brief I would argue that the attribution of purpose is again just a subjective projection on the part of human theorists, and constitutes a potentially misleading gloss on evolutionary processes. The term 'natural selection' can suggest that some sort of choice mechanism is involved, which can in turn suggest a form of proto-intentionality on the part of biology – as if 'Mother Nature' literally chooses the most fit to survive. But of course this is only a metaphorical take on the fact that possessing some aimlessly mutated trait which just happens to constitute an advantage over ones competitors will mechanically *cause* the possessor to propagate more numerously. The actual mechanisms are all straightforwardly causal, and there is no real need to invoke anthropomorphic heuristics appealing to purpose or design. The operational effect of possessing a randomly generated favorable trait will appear *as if* the trait were 'selected', but of course there is no 'invisible hand' at work. It may be an arch conservative stand in contemporary intellectual culture, but I would still concur with Hume that it's a basic conceptual fallacy to try and derive an 'ought' from an 'is'.

2.8 Computational Ascriptions Are Interest Relative

I would now like to propose a different perspective on the issue. Rather than distinguishing 'true' from 'false' cases of implementation, what the various proposed constraints do instead is to go some distance in distinguishing interesting and *pragmatically useful* implementations from the many uninteresting, trivial and useless cases that abound in the space of theoretical possibility. It's certainly true that there is no pragmatic value in most interpretive exercises compatible with ↑MR and ↑MR*. Ascribing computational activity to physical systems is *useful* to us only insofar as it supplies *informative outputs*.

So, interesting and useful mappings are such that we can directly read-off something that *follows from* the implemented formalism, but which we didn't already know in advance and explicitly incorporate into the mapping from the start. That's the incredible value of our computational artefacts, and it's one of the only *practical* motivations for playing the interpretation game in the first place. Hence a crucial difference between our computational artefacts and the attributions of formal structure to naturally occurring open systems, as employed by ↑MR* exercises, is

that the mapping in the latter case is entirely *ex post facto* and thus supplies us with no epistemic gains. The abstract procedural 'trajectory' is already known and used as the basis for interpreting various state transitions in the open system and hence characterizing it as an implementation. In sharp contrast, we can use the intended interpretation of our artefacts both to *predict* their future behaviour, as well as *discover* previously unknown output values automatically.

And this is obviously why an engineered correlation obtains between fine-grained causal structure and abstract formal structure in the case of our artefacts – we want them to be informative and reliable! We also want them to be highly versatile, and this is where counterfactual considerations can come to the fore in practice: over time we do runs on a huge number of different inputs, and in principle the future outputs follow as direct consequences of the intended interpretation. And this is where semantic considerations can enter the picture – the purely syntactic formalisms are designed to *preserve truth* in our intended interpretation, so that from the automated syntactic transformations we can apply our interest-relative semantics and hence discover new truths about our chosen semantic domain. In general, a particular physical device is *useful to us* as a computer only when its salient states are distinguishable by us with our measuring devices, and when we can put the system into a selected initial state to compute the output of our chosen algorithm on a wide range of input values. And these features will be relative to our current technological capabilities.

These *pragmatic* considerations supply clear and well motivated criteria for differentiating useful from useless cases of physical implementation. And I would advocate this type of pragmatic taxonomy in lieu of attempts to give overarching theoretical constraints purporting to distinguish literally 'true' from 'false' cases. The pragmatic factors do not supply global and uniform necessary conditions (and the ever present non-zero probability of error indicates than none is *sufficient*, either). Different desiderata will have shifting roles and prominence in different contexts of application, and will be satisfied to varying *degrees* dependent on the goals and purposes in question, as well as the state of our technological progress. Computation is a highly versatile tool, and there is no single and objective class of phenomena that can be isolated as comprising the 'real' instances of physical implementation. Instead, SMA specifies the maximal and context neutral space of possibilities, and varying pragmatic considerations can then be applied to carve out different subsets within this space which prove useful or interesting according to our divergent human purposes. In short, physical computation is not a natural kind – it is founded upon human convention, interpretation and choice.

2.9 Some Standard Objections

I will end the paper by briefly addressing some objections that often arise in response to this position.

2.9.1 The Spectre of Pancomputationalism

In his excellent and illuminating Stanford Encyclopedia article, Piccinini (2015a) observes that one of the motivations for rejecting SMA is that it induces 'unlimited pancomputationalism', which is presumably something we should wish to avoid. But it's difficult to see why this type of pancomputationalism should constitute a theoretical menace, since it goes hand in hand with anti-realism about physical computation, and simply implies that any number of abstract mappings exist in *a purely mathematical sense*. Analogously, there are any number of abstract mappings that exist from the set of positive integers to collections of physical objects and particles. For example, the set of O_2 molecules in some arbitrarily delimited region of the atmosphere is enumerated via some function on the positive integers. And this region can be defined as a proper subset of some other region and the same molecules are enumerated by any number of different functions. Hence the same molecules can be members of arbitrarily many different sets and images under many different mappings. Is this a threat? For the most part we don't care about all these possible sets and enumerations. But in some cases we do, as in the set of human beings living in some country, when it comes time to do a census.

2.9.2 The Threat to CTM

I endorse a purely formal and non-intrinsic account of computation, and consequently argue that the mathematical theory of computation alone is not sufficient to provide a full explanatory *theory of* particular subject disciplines, such as a *computational theory of the mind*. This is a specialized scientific application that requires many additional resources appropriate to the phenomena and subject area under investigation. Computation is an extremely powerful and versatile formal tool, that can be applied to a virtually limitless range of phenomena. However computation *per se* has no *mystical powers*, and merely implementing the 'right' sort of computational formalism cannot magically transform some given arrangement of mass/energy into a mind. On my account, much more is required than merely implementing a formal procedure. In particular, the system must be able to *do* a host of complex and sophisticated things within a multifaceted environment. See Schweizer (2016) for further discussion.

2.9.3 Not All Levels of Description Are 'Intrinsic' from the Perspective of Physics

There are many levels of description that are not 'intrinsic' from the perspective of fundamental physics, but are nonetheless perfectly legitimate and scientifically

respectable. For example, various arrangements of mass/energy configured in such a way as to perform some clear biological function, such as 'being a kidney'. In response, I would argue that the attribution of computational structure is crucially *disanalogous* to cases such as this, which still trade on characteristics which are themselves essentially physical in nature. In order to be a kidney, a particular assemblage of material stuff must *do things* with other instances of material stuff that are characterized in terms of, e.g. the chemical composition of blood, waste products, filtering, etc. There is an objective, observer independent fact of the matter regarding whether or not a given configuration of matter performs the chemically specified functions required of kidneys, because biological functions are defined in terms of cause and effect relations in the physical world, and in stark contrast, computational realizations are *not*.

There is a pronounced difference here between *actual* versus *abstract* characteristics which makes attributions of computational structure observer dependent in a manner not shared by biological functions. The inputs to a computational system are essentially 'symbolic' rather than physical, where the material implementations of the symbolic or formal inputs must be *interpreted* as such by an outside agent, and where this symbolic interpretation is entirely *conventional* in nature. This marks a prominent discontinuity in levels of description.

2.9.4 There Are Objective Constraints If Given an Appropriate Physical Description

Not just anything goes as SMA seems to suggest – there *are* objective constraints at appropriately specified levels of physical description, e.g. circuit theory (see Scheutz 1999). And I would agree that, relative to particular design parameters imposed by human engineers, in conjunction with known principles of materials science, there can be very tightly constrained abstract solutions. SMA does not imply that such mappings are 'arbitrary', and surely the impressive success and reliability of our artifacts is not a subjective phenomenon. As with Dennett's Intentional Stance, predictive success is an objective criterion. However, to the extent that success *is* achieved, it ultimately rests upon skilled manipulation of the physical substrate. And the ever present possibility of error and malfunction indicates that an abstract computational description of this (continuous) substrate is still a normative idealization and not an 'intrinsic' characterization. There is nothing physically or metaphysically privileged about circuit theory as a level of description, and it does not preclude alternative characterizations and different computational mappings ascribed to the very same physical system. Hence such 'favored' mappings have no impact on the basic SMA perspective.

2.9.5 SMA Cannot Differentiate a Stone from a Sophisticated Computational Artifact

And surely there *is* a difference, objectors will contend, and hence SMA does not provide a satisfactory account of computation in physical systems. To this complaint I would reply that the crucial difference is in our ability to manipulate the artifact in order to acquire new information. Artifacts are specifically designed and built to satisfy non *ex post facto* mappings – this is why they're so useful and why we pay good money for them. But this feature does not ground an ontological distinction between 'real' versus 'spurious' implementations. In other cases we appeal to *ex post facto* methods, as in error checking the very same artifacts. And in the case of 'natural computation', if we have a theory concerning what computation a given biological system is performing, then we can predict future *physical* states of the system, and also test our theory, by carrying out the computation *first* and then looking to see if it maps to the empirical facts.

References

Bishop, J.M. 2009. Why computers can't feel pain. *Minds and Machines* 19: 507–516.

Block, N. 2002. Searle's arguments against cognitive science. In *Views into the Chinese room*, ed. J. Preston and J.M. Bishop. Oxford: Oxford University Press.

Boolos, G., and R.C. Jeffrey. 1989. *Computability and logic*. 3rd ed. Cambridge: Cambridge University Press.

Chalmers, D.J. 1996. Does a rock implement every finite-state automaton? *Synthese* 108: 309–333.

Chrisley, R.L. 1994. Why everything doesn't realize every computation. *Minds and Machines* 4: 403–420.

Copeland, J. 1996. What is computation? *Synthese* 108: 335–359.

Dennett, D. 1981. True believers: the intentional strategy and why it works. In A. F. Heath (Ed.) *Scientific Explanation: Papers Based on Herbert Spencer Lectures given in the University of Oxford*, Oxford: University Press.

Fodor, J. 1981. The mind-body problem. *Scientific American* 24: 114.

Kripke, S. 1982. *Wittgenstein on rules and private language*. Cambridge: Harvard University Press.

Maudlin, T. 1989. Computation and consciousness. *Journal of Philosophy* 86 (8): 407–432.

Milkowski, M. 2013. *Explaining the computational mind*. Cambridge: MIT Press.

Newman, M. 1928. Mr. Russell's "Causal Theory of Perception". *Mind* 37: 137–148.

Piccinini, G. 2015a. Computation in physical systems. In *The Stanford encyclopedia of philosophy,* ed. E.N. Zalta. http://plato.stanford.edu/archives/fall2015/entries/computation-physicalsystems/.

———. 2015b. *Physical computation*. Oxford: Oxford University Press.

Putnam, H. 1988. *Representation and reality*. Cambridge: MIT Press.

Rescorla, M. 2014. A theory of computational implementation. *Synthese* 191: 1277–1307.

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

Schweizer, P. 2012. Physical instantiation and the propositional attitudes. *Cognitive Computation* 4: 226–235.

———. 2016. In what sense does the brain compute? In *Computing and philosophy*, Synthese library 375, ed. V.C. Müller, 63–79. Heidelberg: Springer.

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

Shagrir, O. 2001. Content, computation and externalism. *Mind* 110 (438): 369–400.

Sprevak, M. 2010. Computation, individuation, and the received view on representations. *Studies in History and Philosophy of Science* 41: 260–270.

Turing, A. 1936. On computable numbers, with an application to the entscheidungsproblem. *Proceeding of the London Mathematical Society*, (series 2) 42: 230–265.

———. 1950. Computing machinery and intelligence. *Mind* 59: 433–460.

Chapter 3
The Notion of 'Information': Enlightening or Forming?

Francois Oberholzer and Stefan Gruner

Was der Philosoph schreibt, ist für den Informatiker nur zum geringen Teil akzeptabel, und umgekehrt.
— HEINZ ZEMANEK

Abstract 'Information' is a fundamental notion in the field of artificial intelligence including various sub-disciplines such as cybernetics, artificial life, robotics, etc. Practically the notion is often taken for granted and used naively in an unclarified and philosophically unreflected manner, whilst philosophical attempts at clarifying 'information' have not yet found much consensus within the science-philosophical community. One particularly notorious example of this lack of consensus is the recent Fetzer-Floridi dispute about what is 'information'—a dispute which has remained basically unsettled until today in spite of a sequence of follow-up publications on this topic. In this chapter our philosophical analysis reveals with reference to Gottlob Frege's classical semiotics that the above-mentioned Fetzer-Floridi dispute cannot come to any solution at all, because the two competing notions of 'information' in that dispute are basically synonyms of what Frege had called 'sense' (*Sinn*) versus what Frege had called 'meaning' (*Bedeutung*). As Frege had convincingly distinguished sense and meaning very clearly from each other, it is obvious that 'information' understood like 'sense' and 'information' understood like 'meaning' are incompatible and cannot be reconciled with each other. Moreover we also hint in this chapter at the often-forgotten pragmatic aspects of 'information' which is to say that 'information' can always only be 'information *for somebody*' with regard to a specific aim or goal or purpose. 'Information', such understood, is thus a *teleological* notion with a context-sensitive embedding into what the late Wittgenstein had called a 'language-game' (*Sprachspiel*). Shannon's quantified notion of 'information', by contrast, which measures an amount of

F. Oberholzer and S. Gruner (✉)
Department of Computer Science, University of Pretoria, Pretoria, Republic of South Africa
e-mail: sg@cs.up.ac.za

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_3

unexpected surprise and which is closely related to the number of definite yes-no-questions which must be asked in order to obtain the desired solution of a given quiz puzzle, is *not* the topic of this chapter—although also in Shannon's understanding of 'information' the quiz puzzle scenario, within which those yes-no-questions are asked and counted, is obviously purpose-driven and Sprachspiel-dependent. We conclude our information-philosophical analysis with some remarks about which notion of 'information' seems particularly amenable and suitable for an autonomic mobile robotics project which one of the two co-authors is planning for future work. To separate this suitable notion of 'information' from other ones a new word, namely ≪enlightation≫, is coined and introduced.

Keywords Philosophy of information · Data · Sense · Meaning · Structure · Purpose

3.1 Introduction

Many researchers in the field of artificial intelligence (AI) aim at creating instances of 'strong' general AI including autonomic, dynamic and self-adaptive problem solving behaviour in a-priori uncharted and possibly changeable environments. In such scenarios we can imagine a robot with sensors that can perceive data about the world around it, interpret such data as relevant 'information', process this information further by various methods of reasoning, and make action-oriented decisions based on the results of such reasoning. Such a robot could thus be called an 'information processor'.

Alas the very concept of 'information' is still in need of science-philosophical clarification nowadays—several decades after the intuitive and often fashionable coining of discipline-defining buzzwords such as 'information theory', 'information science', 'information technology', 'information systems', 'information processing', 'informatics', and the like (Aspray 1985; Geoghegan 2008; Kline 2006). Until today there are many conflicting definitions and science-philosophical disputes about what can or what cannot be characterised as 'information' (Floridi 2016, pp. 2-3). Early information-philosophical considerations can already be found in various writings by the Austrian computer pioneer Heinz Zemanek (Gruner 2016), whilst a recent, comprehensive, Wittgenstein-influenced overview-essay on the question of what is 'information' beyond a mere definition can be found in Böll and Cecek (2015) where many different semiotic facets and aspects of information are described in much detail. Also according to Claude Shannon the word 'information' has been given different meanings by different authors already within the rather narrowly defined discipline of information theory, whereby it is not to be expected that one single concept of 'information' would satisfactorily account for the numerous possible applications (Floridi 2005). Thus, as far as our envisaged robot application is concerned, we have to ask ourselves: what really is this something that our robot would take in, process and base its decisions on?

An interesting debate on this topic—i.e.: the notion of 'information'—took place during the years 2004–2005 between Luciano Floridi and James Fetzer. Even now, more than a decade later, a number of lessons can be learned from that academic dispute, which eventually ceased without consensus and was not mentioned in the recent overview-essay of Böll and Cecek (2015). In this chapter we will recapitulate the most important points of that debate and show that it was based on talking about two quite different concepts under the guise of the same name. That debate started with Floridi's outline of his theory of so-called 'strongly semantic' information (Floridi 2004b) wherein he introduced a new quantitative information theory to solve the so-called Bar-Hillel-Carnap paradox.[1] Fetzer objected to one aspect of that theory, namely that it requires information to be true to count as 'information' (Fetzer 2004), whereupon Floridi responded in Floridi (2005) with numerous arguments in order to support his own claim that something is 'information' only if it is true. In the subsequent sections will recapitulate those three papers, as well as other immediately relevant literature, and try to assess the plausibility and suitability of their claims, particularly with regard to our above-mentioned robotic scenario.

3.2 Strongly Semantic Information and the Fetzer-Floridi Dispute

The Bar-Hillel-Carnap (BHC) paradox (Salmon 2003) refers to a strange situation in which an inconsistent (self-contradictory) sentence, which no reasonable receiver would accept, is regarded as carrying more 'semantic information' than a contingently true sentence (Floridi 2004b), i.e.: a proposition which is logically satisfiable and empirically satisfied however not a trivial formal tautology. With his theory of strongly semantic information (TSSI), Floridi dissolved the BHC paradox, albeit at the cost of sacrificing the classical notion according to which information consists of structured data—see Zemanek for comparison (Gruner 2016)—which might not necessarily be 'true':

[1]The question might arise at this point whether or not all information is per-se semantic (i.e.: meaningful), such that the term 'semantic information' (Floridi 2016, pp. 44–49) would be a pleonasm, coming close to a tautologism like 'wooden wood'? Due to the inherent grammatical weakness of the English language, the term 'semantic information' is ambiguous and could be interpreted either as 'information *with* semantics', which seems to be the above-mentioned pleonasm, or as 'information *about* semantics', i.e.: some kind of meta-information concerning purely theoretical-linguistic entities—as opposed to, for example, 'information about birds' or 'information about health'. For further details see the Semantic Concepts of Information, online at http://plato.stanford.edu/entries/information-semantic/ in the Stanford Encyclopedia of Philosophy. For recent comments about the Bar-Hillel-Carnap paradox (presented at a reasonably high level of mathematical formality) the reader might want to look at a technical report (Gorsky and Carnielli 2013) which is available online, too.

> The main hypothesis supported has been that semantic information encapsulates truth, and hence that false information fails to qualify as information at all (Floridi 2004b, p. 25).

Although it seems as if Floridi wanted to replace the classical theory with his TSSI, we are reassured by Scarantino and Piccinini (2010) that Floridi holds a 'non-reductionist' position according to which information as an 'explicandum' can be explained in a variety of ways, depending on the various meta-theoretical viewpoints and requirements under which a theory of information could possibly be conceived; for comparison see (Floridi 2016, p. 2),[2] Floridi's overview article (Floridi 2004a), as well as his online-entry on semantic concepts of information in the Stanford Encyclopedia of Philosophy, according to which 'semantic information' can be either 'factual' or 'instructional', although the 'instructional' aspect is often left out. This 'factual' type of 'semantic information' is closely associated with the notion of 'knowledge' from an epistemological or gnoseological point of view, such that information as 'true semantic content' appears to be a necessary condition for knowledge.[3]

When contemplating the 'essence' of the TSSI, the modern mind will almost automatically raise the question: how do we recognise what is truly true? Or: would we live in an information-less world if we would have to concede that absolute truth is not for us to be possessed in this world? Hence: what is the very notion of 'truth' which is tacitly presupposed underneath a philosophem such as the TSSI? Moreover: would we not fallaciously attempt to 'explain' one mystery, namely: what is 'information', by means of an even more mysterious mystery, namely: what is 'truth'? Such is indeed the core of Fetzer's argument against the TSSI. We see in all modern philosophy of science that what is regarded as empirically true is historically variable—which has in combination with the TSSI the practically odd consequence that only ex-post-facto, in hindsight, we would be able to decide whether or not we had received some 'information' in the past. Consequently—and against our intuition—entire textbooks for generations of students in higher education could ultimately turn out to be completely 'information'-less after the empirical sciences have made sufficient progress. Fetzer himself provided the following argument:

> If we encounter the sentence 'there is life elsewhere in the universe', ostensibly in English, we, as speakers of English, would find it meaningful data, but we would not know whether it is true. On the standard conception, that sentence would qualify as information that may be false. On Floridi's conception—even as meaningful data—it might or might not count as information, since it properly qualifies as 'information' only if it is true (Fetzer 2004, p. 225). It follows that, on Floridi's account, a sentence that is information can have a negation that is not, where no one knows which is information and which is not! This result must be at least as paradoxical as any it would resolve (Fetzer 2004, p. 226).

[2] "*Non-reductionists like myself...*"

[3] At this point the reader might remember some classical works by Plato according to whom knowledge must necessarily be true or otherwise it would not be 'knowledge': false knowledge, for Plato, would be a meaningless contradiction in terms. Modern epistemologies, by contrast—such as, for example, Popper's—have loosened the hitherto tight connection between the concepts of 'knowledge' and 'truth'.

In addition to Fetzer, and from our specific perspective as computer scientists, we must also ask if we could ever 'inform' (in-*form*) a computer—which does not have any conscious notion of 'truth' at all—by means of some control-input if Floridi's TSSI is a suitable and appropriate philosophem. Is the very word ≪information≫ then no longer an acceptably speakable word in this context? Another example scenario: a cat and a rabbit, both sitting calmly on the grass, suddenly see a mouse emerging from its mouse-hole. Upon seeing this same mouse, the cat will suddenly change its behaviour quite dramatically, whilst the rabbit will remain rather unmoved by the same visible phenomenon. However, a mouse does not possess any Boolean 'truth value' at all—a mouse is not a propositional sentence. Thus: neither is the mouse 'true' for the cat, nor is the mouse 'false' for the rabbit. Are we then, according to the TSSI, no longer allowed to say in this *Sprachspiel* situation that the cat's behaviour—unlike the rabbit's—was 'informed' (in-*formed*) by the sudden appearance of the mouse? Moreover—as much or little as what we know about the minds of cats—the cat is also not able to project its observation onto a linguistic-logical meta level in the form of an assertion of the kind: "the proposition 'the mouse is delicious food for me' is true". Thus, within the philosophical framework of Floridi's TSSI, the appearance of the mouse cannot have 'informed' the cat about anything—in contrast to the obvious *causal* relation between the mouse's observable appearance and the cat's equally observable change of attitude and behaviour. In other words: Floridi's notion of 'information' is devoid of any relationship with the notion of 'causation' although in our field of informatics (including sub-disciplines such as AI and cybernetics) those two notions are indeed quite closely related to each other; for comparison see Illari and Russo in Floridi (2016, pp. 235–248). With regard to the question of whether or not 'environmental information' can be regarded as 'semantic' the reader is referred to Scarantino and Piccinini (2010, p. 314).

In defence of Floridi it was pointed out by Sequoiah (2007) that the plausibility of the TSSI can be strengthened if its applicability is explicitly restricted to the declarative objective semantic domain (DOS). In particular: if some statement (proposition) is objectively true, it would be information per-se, regardless of whether or not anyone knows that it was true. Regardless of any particular agent,

> DOS information's status as information is independent of epistemic access (Sequoiah 2007, p. 13).

The intellectual proximity of such assertions to the philosophical position of 'Platonism' (with its characteristic notion of time-less ideas) is evident. Floridi himself listed in Floridi (2005, sect. 4),

> nine bad reasons to think that false information is a type of semantic information

in order to defend his position that information has to be true to count as 'information'. Alas the crucial term 'semantic information' was not given a precise lexical definition, such that it remains somewhat difficult to grasp what was actually being contested and defended in those disputes. Anyway, as propositional sentences are the *only* entities in our universe of discourse which could possibly be 'true'

or 'false'—neither mathematical objects such as numbers nor natural entities such as electromagnetic waves have this possibility—it is evident that Floridi, by the use of phrases like 'true', 'false', 'false information', and the like, has already presupposed that any 'information' whatsoever must possess the linguistic-logical form of a proposition. Floridi's could a thus be called a position of 'Linguicism'—as opposed to 'Naturalism' or 'Physicalism'—within the wider area of the information-philosophical discourse, and all the finer details of the TSSI-DOS-based theories of information are then being debated within the perimeters of such 'Linguicism'. Evidence of such 'Linguicism' can indeed be found in statements such as this:

> The new version of the definition (RSDI) now describes DOS information as well-formed, meaningful and truthful data (Floridi 2005, sect. 7),

which, by the way, tacitly conflates the notion of 'data' with the notion of 'logical proposition' by attributing 'truth' to mere data (which are typically truth-less entities like mathematical numbers). Comparing this viewpoint to the way in which *five* kinds of 'information' were distinguished by Zemanek (1970, sect. 3), namely 'numerical', 'physical', 'formatted', 'natural text', and 'formal text', we can see that Floridi had limited himself to a declarative form of the latter two.

Strictly speaking, Floridi in his role as a TSSI-DOS-philosopher could not even accept tonight's televised weather forecast for tomorrow as TSSI-DOS 'information' because only tomorrow we will know whether or not tonight's forecast proposition was true—in contrast to all our daily life experience wherein millions of ordinary people are indeed regarding tonight's televised weather forecast for tomorrow as highly valuable 'information', upon which they make numerous practical preparations, although they do not know its truth at that point in time; see Fig. 3.1 for an illustration. Regardless of whether true or not, the weather forecast for tomorrow in-*forms* the people by *forming* new ideas in their minds. The same can be said for the entire discourse in the empirical sciences, in which almost all scientific statements (such as the weather forecast for tomorrow) are *hypothetical*, not propositional (Bunge 1998, ch. 5). Hence, according to the DOS-TSSI in its strict interpretation, the entire scientific discourse including millions of scholarly letters would be almost devoid of any information due to its intrinsically hypothetical character—in contrast to thousands of scientists' perceived experience of hypotheses as informative. Newton's classical law of gravity, for example, is—strictly speaking—false (as we know today): does then a physics teacher at school not 'inform' his pupils when he is teaching Newton's classical law of gravity to them?

In continuation of the same discourse, Scarantino and Piccinini (2010) published their paper on the topic of information without truth, in which they critiqued Floridi's veridicality thesis; Scarantino and Piccinini (2010) distinguished between 'natural information'—i.e.: physical signs which can be regarded as symptoms of some event or system state—and 'non-natural information'—i.e.: intentional signs which are meant to carry purposeful messages usually by virtue of convention—such as, for example, three rings of a bell signifying that a commuter bus is fully occupied and ready for departure. As mentioned above, Floridi's philosophy of

Fig. 3.1 Does this weather forecast map contain or provide information, and—if yes—can the TSSI-DOS theory deal with it?

information belongs into 'non-natural' realm in terms of Scarantino and Piccinini (2010). With regard to the notion of 'false information' as discussed in Floridi (2005), Scarantino and Piccinini presented examples of composite propositional sentences which are false and pass the so-called 'splitting test' of attributive versus predicative usage of grammatical adjectives such as 'nice', 'good', 'true', and the like—although Floridi had argued in Floridi (2005) that the term 'false information' (in his terminology) would be characterised by *failing* the splitting test. Note that the lexical *word* ≪false≫ in that context appears as a grammatical meta-predicate like in 'false gold' or a 'false friend'—not as a Boolean semantic truth value of a given proposition. This whole debate about the term 'false information' is very subtle and somewhat error-prone, since it hinges on the apparently self-referential 'natural semantics' of the meta-predicate *word* ≪false≫ which—unlike, for example, the predicate word ≪red≫—seems to signify itself as its own meaning. In those kind of philosophical arguments, in which ≪false≫ is a grammatical meta-predicate word and 'false' is a Boolean truth value, the level differences between object language and meta language can easily get confused.

Indeed: *if* FALSE INFORMATION is (for Floridi) an *ontic* impossibility (such as, for example, DEHYDRATED WATER which cannot exist), then the lexical *word* ≪false≫ in the conceptual term 'false information' *must* be understood in analogy to the word ≪false≫ in the term 'false gold', which superficially appears to be gold though it is not. There is, however, something which gold and false gold have apparently in common, namely its yellow-metallic surface; otherwise the term 'false

gold' would not make any sense in the *Sprachspiel* of a community of speakers who know what GOLD is. By analogy, also the term 'false information' can only make sense in a community of speakers if it somehow resembles and has something in common with information, such that the cross-reference is making sense. In other words: *if* FALSE INFORMATION is an *ontic* impossibility, to which the term 'false information' can thus *not refer*, then the term 'false information' can *either not make any sense* at all in the *Sprachspiel* of our community of speakers, *or the term must refer to something else* which possesses the superficial appearance of information, whereby this superficially common property amongst the two is nothing else but the empty shell of a formal-logical proposition in its lexical-grammatical expression—hence our earlier characterisation of Floridi's information-philosophical position as 'Linguicism'. From the vantage point of such 'Linguicism' it is, in particular, not possible to 'see' that the appearance of false gold in the display-window of a jewellery shop can equally well as genuine gold motivate a jewellery collector to enter this shop and to make a purchase. Anyway:

> Whether false information passes the splitting test depends on whether we accept that a false p can constitute information,

said Scarantino and Piccinini (2010, p. 321). With regard to Floridi's later argument about the 'semantic loss' through modifications of text (Floridi 2007), Scarantino and Piccinini noted that such argument would be based on thinking about information as something quantitative rather than qualitative:

> When we worry about information loss, we are not primarily—if at all—concerned with the quantity of information contained in a repository. Rather, we are generally concerned with whether an information repository carries the same information it originally carried (Scarantino and Piccinini 2010, p. 322).

In that context Scarantino and Piccinini also mentioned an example in which every true proposition in a chemistry book is replaced with a true proposition from a biology book. According to Scarantino and Piccinini no informational loss would occur in such a scenario in a merely quantitative sense, though a major loss has been suffered in the qualitative sense since the modified book's message is no longer the same as before (Scarantino and Piccinini 2010), or, as we could say more precisely in Gottlob Frege's terms: only the book's sense (*Sinn*) has been altered, whilst its meaning (*Bedeutung*)—namely the set of Boolean truth values of its sentences—has survived the modification.

Moreover, in many practical situations we are dealing with what Scarantino and Piccinini have called 'non-truth-evaluable information' (Scarantino and Piccinini 2010), for which there is no room in Floridi's philosophical framework. As far as our own discipline of *computer science* (including: informatics, AI, cybernetics, and the like) is concerned, into which our information-philosophical papers considerations are purposefully embedded, Scarantino and Piccinini stated correctly:

> Computer scientists routinely label as 'information processing' all the cases in which computers process semantically evaluable structures, whether they are true or false (Scarantino and Piccinini 2010, p. 324).

For comparison see Zemanek's early computer-philosophical papers concerning various philosophical aspects of automated information processing (Gruner 2016), in which he had made quite similar remarks already several decades ago:

> If the computer could do nothing but book-keeping and solving equations, the situation would be much easier. But the computer can do much more, as I will show, as you all know; the computer processes all kinds of information which is then used for decisions on both the technical and the human level; the computer can be looped into flows of information and decisions which are predominantly on the human level—but unlike the human computer the electronic computer cannot switch its thinking from the logical to the human level (Zemanek 1974, pp. 899–900).

Hence, as Scarantino and Piccinini plausibly concluded: if Floridi's radical notion of 'information' as 'truth' were to be generally accepted in all domains, then the entire computer science community would suddenly stand on the wrong side of the philosophical fence and would be forced to develop a cumbersome philosophical theory of 'automated *non*-information processing'. In hindsight it seems that Zemanek, one of the most important founding-fathers of computer-philosophy, was indeed visionary when he wrote more than four decades ago:

> What the philosopher writes is hardly acceptable for the informatician, and vice versa[4] (Zemanek 1973, p. 385).

Floridi might then perhaps argue in return—as we have to concede—that we could in this case simply and better (and more modestly) speak merely of 'automated data processing', (instead of 'automated information processing'), because most of the data inside the computer's storage can be 'information' *only for us* as the computer's purposeful users—not for the intentionally indifferent machine itself.

3.3 Enlightening versus Forming

After having analysed all those arguments, particularly the ones exchanged between Fetzer and Floridi, it seems to us that the foregoing dispute is not a case of a debate in search of the best univoque definition of a single concept—see again (Böll and Cecek 2015) for comparison—but rather a 'clash' of incommensurable concepts under the guise of merely the same *word*: ≪information≫.[5] For Floridi and Sequoiah-Grayson, as we have seen, TSSI-DOS information is a collection of some rather abstract, so-to-say 'neo-Platonic' sentences which objectively describe a rather static-ideal world independent of any agent and its purposeful intentions—somewhat similar to the rigid world of the early Wittgenstein's *Tractatus Logico-*

[4] In the German original: " *Was der Philosoph schreibt, ist für den Informatiker nur zum geringen Teil akzeptabel und umgekehrt.*"

[5] Similar terminological-conceptual confusions have become notorious also other sub-disciplines of informatics, particularly in *software architecture* (Gruner 2014), as well as in *digital forensics* (Tewelde et al. 2015).

Philosophicus as analysed already four decades ago by our pioneer Heinz Zemanek (Gruner 2016).

> The Tractatus is in fact a synthesis of the theory of propositional calculus and the theory of the sentence as a picture of reality. One of Wittgenstein's basic assumptions is that there are elementary sentences, atom-like statements, which he calls elementary propositions and which are logically and factually either true or false (Zemanek 1975, p. 24).

These elementary propositions taken together can then give us a perfect description of the world. Fetzer, Scarantino and Piccorini, on the contrary, have characterised information as something practical, something that intentional and context-sensitive beings use to understand their world, to make decisions, to communicate. For those authors, information *emerges* always-only as information *for* someone, not per-se, and it is mostly something which *triggers effects* when it is somehow 'understood' in its specific situation.

At this point we ought to ask ourselves: should we simply 'give up' and accept such incommensurable coexistence of concepts? Or should we still strive to synthesize existing theories of information—a.k.a. information-philosophical 'reductionism' (Floridi 2016, p. 2)—and still attempt to redefine our most fundamental terms as appropriately as possible, such as to eliminate equivocity as best as we can for the sake of scientific unity? For comparison the reader might remember—for example—the semantic history of the fundamental terms '*force*' and '*energy*', on which the discipline of physics is based: *if* our discipline of informatics (including its various sub-disciplines) ought to be based on a few fundamental concepts like 'information'—such as physics on 'force' and 'energy'—then the internal unity and theoretical coherence of our scientific discipline will stand or fall with a common (communal) understanding of the term 'information' in a similar manner in which the intra-disciplinary coherence and unity of physics depends on the common understanding of 'force' and 'energy' among the *Sprachspiel* community of physicists. Even during the most shattering paradigmatic crisis in the history of modern physics, namely the quarrel between Albert Einstein and Niels Bohr about the right interpretation of quantum theory, their notion of 'energy' was shared as the same. The historic-semantic variability in the old notion '*aether*'—by contrast—was never perceived as problematic, because 'aether' was never a fundamental pillar concept in the innermost theoretical core of the discipline (Gruner and Bartelmann 2015).

Would it, thus, not be better to distinguish those incommensurably different concepts of 'information' by different lexical terms, and thus approach some theoretically desirable univocity, instead of continuing to use the same *word* ≪information≫ equivocally for all of them—and thereby also concede that all those notions of 'information' continue to have their own 'right to exist' as long as they cannot be confused with each other anymore? I.e.: we ought to need one name for the objective descriptions of the world as per Floridi (et al.), and another name for what stimulated brains and computers process and base their context-dependent and purpose-directed decisions on.

In the Dutch-based South-African language *Afrikaans* we have indeed *two* suitable words in this semantic field of discourse: ≪inligting≫ and ≪informasie≫. Though they are widely regarded as synonyms, etymologists will easily recognise that ≪inligting≫ stems from the *Germanic* root for 'light'—i.e.: to illuminate, enlighten, be enlightened or reach enlightenment—whereas ≪informasie≫ stems from the *Romanic* root for 'form' or 'shape'. Thus we might speculate philosophically at this point that the notion of reaching light and becoming enlightened of the truth is closer to Floridi's quasi-Platonic concept of TSSI-DOS 'information', whereas forming an idea or image in someone's mind, even a wrong or inappropriate image, is closer to Fetzer's notion of stimulative-effective 'information' as it is commonly understood in our technological area of informatics, cybernetics, and various related disciplines.

Mapping this terminological distinction from Afrikaans to English for wider usage, we herewith suggest that Fetzer's notion of 'information' shall still continue to be named ≪information≫, as this is the standard or common usage of the word in our technical-academic discipline in good agreement with what etymology is teaching us. In addition, we also introduce the following completely *new word* to denote the concept of 'information' which Floridi has described; this might henceforth be named ≪*enlightation*≫. By dividing information from enlightation in this way, we hope to be able end the foregoing information-philosophical dispute, such as to make some progress towards solid terminological foundations for an intra-disciplinary coherent science of informatics. Of course much debate will still arise about these definitions, but at least the 'veridical nature' of 'information' should no longer trouble us anymore.

3.4 Conclusion and Outlook

On the basis of our philosophical analysis of the notorious Fetzer-Floridi dispute about the characteristics of 'information' we have coined and proposed in this chapter a *new term*, namely '*enlightation*', with the hope of closing the debate on the 'veridical nature' of information. From a more practically and AI-oriented point of view we had to ask how our (envisaged) autonomic robot might perceive (and survive in) its environment, and hence what it is that can rightly be called 'information' *for* such a robot—i.e.: under proper consideration of information's pragmatic and teleological aspects, without which information would not be 'information'. Other and more than mere raw *data*, information always *emerges* as something which is *interpreted* in a *context-sensitive Sprachspiel*—i.e.: as information *for* somebody with regard to this somebody's intents, aims, goals, and purposes. In other words: the very same data D can be regarded as 'informative' by some sensitive and sensible entity, X, and as 'not informative' by some other sensitive and sensible entity, Y. If we now ask, again, what it is that our robot will take in, process and base its decisions on, we can answer at this point that it will be 'information' in the way in which Fetzer, Scarantino and Piccinini have used and understood the

term. This is *not* to state apodictically that Floridi's notion of 'information' would be anyhow 'wrong' as such—it is just not the appropriate notion in the context of our robotic scenario, as we have shown. Digital images from electronic camera eyes, audio signals from microphones, chemical traces influencing an artificial nose or an artificial tongue, as well as haptic feedback from surface sensors are the phenomena by which the artificial brain of our robot will be *stimulated*—i.e.: 'in*formed*'. Indeed, a subtle echo of the ancient Aristotelian notion of *causa formalis* can still be recognised in this life-oriented notion of 'information'. In context of the given examples we can see clearly that a life-'informing' stimulus does *not* already have 'semantic' qualities per-se: such a stimulus can either be present or absent, however it can neither be true nor false. What can indeed be true or false are *sentences* (logical judgments) *about* the perceived or imagined phenomena; however it would be an *ontological category flaw* to conflate a sentence *about* some S with this S itself. A stimulus is not a sentence. Hence we must also not conflate the phenomenon of becoming and being informed (i.e.: becoming stimulated and influenced through the interpreted perception of data as 'relevant for me' or as 'not relevant for me') with inter-subjectively truth-able sentences (logical judgments) *about* such a becoming at the linguistic meta-level. Consequently it is well possible to become or to be wrongly informed—such as in the sad example of an autopilot crashing an aircraft against a cloud-covered mountain because of a defective altitude meter amongst the machine's set of instruments. In other words: in the AI-robotic context, by which our chapter was motivated, truth is *not* constitutive for the essential 'information-*ness*' of information, whereas the possible 'falsity' of information is a secondary matter at yet another higher meta-level of interpretation and linguistic *representation of* information. As Gottlob Frege's classical theory of semiotics had made a clear and plausible distinction between a sentence's *intensional* sense (*Sinn*), which is truth-free, and the same sentence's *logical* meaning (*Bedeutung*), which is a Boolean truth-value in $\{T, F\}$, we may thus say that our robot's appropriate notion of 'information' is more closely related to Frege's notion of *Sinn* than to Frege's notion of *Bedeutung*—whereas our newly identified 'enlightation' is in its truthfulness closely related to Frege's *Bedeutung*. An investigation into the most appropriate forms and structures for the efficient *representation* of information—so understood—in the robot's artificial brain remains a future work project for one of the co-authors of this chapter. The *formal-structural aspects* of information, which make their essential contributions to its very information-*ness* according to Zemanek (Gruner 2016), as well as many other aspects and facets according to Böll and Cecek (2015) and Floridi (2016), must thereby also be taken into account.

Acknowledgements Many thanks to *James Fetzer* for an interesting exchange of e-mails on the topic of this chapter some time ago. Many thanks to the anonymous reviewers for their constructive remarks before the presentation of our work at the IACAP'2016 conference in Ferrara (Italy) in June 2016. Many thanks also to the philosophical society '*Footnotes to Plato*' at the University of Pretoria for the opportunity to present our work to them, and for their interesting and insightful feedback.

References

Aspray, W. 1985. The scientific conceptualization of information: A survey. *IEEE Annals of the History of Computing* 7(2): 117–140.

Böll, S., and D. Cecek-Kecmanovic. 2015. What is 'Information' beyond a definition? In Proceedings 36th ICIS, Paper #1363, Fort Worth.

Bunge, M. 1998. *Philosophy of Science*. *From Problem to Theory*, vol. 1, Rev. ed. New Brunswich: Transaction Publ.

Fetzer, J. 2014. Information: Does it have to be true? *Minds and Machines* 14(2): 223–229.

Floridi, L. 2004a. Information. In *The Blackwell Guide to the Philosophy of Computing and Information*, ed. L. Floridi, 40–62. Malden: Blackwell Publishing.

Floridi, L. 2004b. Outline of a theory of strongly semantic information. *Minds and Machines* 14(2): 197–222.

Floridi, L. 2005. Is information meaningful data? *Philosophy and Phenomenological Research* 70(2): 351–370.

Floridi, L. 2007. In defence of the veridical nature of semantic information. *European Journal of Analytic Philosophy* 3(1): 31–42.

Floridi, L. (ed.) 2016. *The Routledge Handbook of Philosophy of Information*. New York: Routledge Publishing.

Geoghegan, B. 2008. The historiographic conceptualization of information: A critical survey. *IEEE Annals of the History of Computing* 30(1): 66–81.

Gorsky, S., and W. Carnielli. 2013. Information, Contradiction, and the Bar-Hillel-Carnap Paradox. Technical Report GLTA-CLE e IFCH, Universidade Estadual de Campinas, Brasil.

Gruner, S. 2014. On the historical semantics of the notion of 'software architecture'. *Journal for Transdisciplinary Research in Southern Africa* 10(1): 37–66.

Gruner, S. 2016. Heinz Zemanek's Almost Forgotten Contributions to the Early Philosophy of Informatics. Paper 1 in Proceedings ACIS'2016: 27th Australasian Conference on Information Systems, Wollongong, Dec 2016.

Gruner, S., and M. Bartelmann. 2015. The notion of 'Aether': Hegel versus contemporary physics. *Cosmos and History* 11(1): 41–68.

Kline, R. 2006. Cybernetics, management science, and technology policy: The emergence of 'information technology' as a keyword 1948–1985. *Technology and Culture* 47(3): 513–535.

Salmon, N. 2003. Reference and Information Content: Names and Descriptions. In *Handbook of Philosophical Logic*, vol. 10, ed. D. Gabbay and F. Guenther, 39–85. Dordrecht: Springer.

Scarantino, A., and G. Piccinini. 2010. Information without truth. *Metaphilosophy* 41(3): 314–330.

Sequoiah-Grayson, S. 2007. The metaphilosophy of information. *Minds and Machines* 17(3): 331–344.

Tewelde S., S. Gruner, and M. Olivier. 2015. Notions of 'Hypothesis' in Digital Forensics. In *Advances in Digital Forensics*, vol. XI, ed. Gilbert Peterson and Sujeet Shenoi, 29–43. Cham: Springer.

Zemanek, H. 1970. Some Philosophical Aspects of Information Processing. In *Proceedings of the 10th Anniversary Celebration of the IFIP: The Skyline of Information Processing*, 93–140. Amsterdam: North-Holland Publishing.

Zemanek, H. 1973. Philosphie der Informationsverarbeitung. *Nachrichtentechnische Zeitschrift* 26(8): 384–389.

Zemanek, H. 1974. The Computer: A Mechanical Device in a Live Environment. In *Proceedings 6th Australian Computer Conference*, Sidney, 894–911.

Zemanek, H. 1975. The Human Being and the Automaton. In *Proceedings of the IFIP Conference on Human Choice and Computers: Human Choice and Computers*, 3–30. Amsterdam: North-Holland Publishing.

Part II
Logic

Chapter 4
Modal Ω-Logic: Automata, Neo-Logicism, and Set-Theoretic Realism

David Elohim

Abstract This essay examines the philosophical significance of Ω-logic in Zermelo-Fraenkel set theory with choice (ZFC). The dual isomorphism between algebra and coalgebra permits Boolean-valued algebraic models of ZFC to be interpreted as coalgebras. The modal profile of Ω-logical validity can then be countenanced within a coalgebraic logic, and Ω-logical validity can be defined via deterministic automata. I argue that the philosophical significance of the foregoing is two-fold. First, because the epistemic and modal profiles of Ω-logical validity correspond to those of second-order logical consequence, Ω-logical validity is genuinely logical, and thus vindicates a neo-logicist conception of mathematical truth in the set-theoretic multiverse. Second, the foregoing provides a modal-computational account of the interpretation of mathematical vocabulary, adducing in favor of a realist conception of the cumulative hierarchy of sets.

Keywords Modal Ω-logic · Ω-logical Validity · Modal Coalgebraic Automata · Neo-Logicism · Set-theoretic Realism

4.1 Introduction

This essay examines the philosophical significance of the consequence relation defined in the Ω-logic for set-theoretic languages. I argue that, as with second-order logic, the modal profile of validity in Ω-Logic enables the property to be

The original version of the chapter has been revised. A correction to this chapter can be found at https://doi.org/10.1007/978-3-030-01800-9_23.

Forthcoming in the 'Proceedings of the 2016 Meeting of the International Association for Computing and Philosophy'.

D. Elohim (✉)
Arché Philosophical Research Centre, University of St Andrews, St. Andrews, Scotland

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_4

epistemically tractable. Because of the dual isomorphism between algebras and coalgebras, Boolean-valued models of set theory can be interpreted as coalgebras. In Sect. 4.2, I demonstrate how the modal profile of Ω-logical validity can be countenanced within a coalgebraic logic, and how Ω-logical validity can further be defined via automata. In Sect. 4.3, I examine how models of epistemic modal algebras to which modal coalgebraic automata are dually isomorphic are availed of in the computational theory of mind. Finally, in Sect. 4.4, the philosophical significance of the characterization of the modal profile of Ω-logical validity for the philosophy of mathematics is examined. I argue (i) that it vindicates a type of neo-logicism with regard to mathematical truth in the set-theoretic multiverse, and (ii) that it provides a modal and computational account of formal grasp of the concept of 'set', adducing in favor of a realist conception of the cumulative hierarchy of sets. Section 4.5 provides concluding remarks.

4.2 Definitions

In this section, I define the axioms of Zermelo-Fraenkel set theory with choice. I define the mathematical properties of the large cardinal axioms to which ZFC can be adjoined, and I provide a detailed characterization of the properties of Ω-logic for ZFC. Because Boolean-valued algebraic models of Ω-logic are dually isomorphic to coalgebras, a category of coalgebraic logic is then characterized which models both modal logic and deterministic automata. Modal coalgebraic models of automata are then argued to provide a precise characterization of the modal and computational profiles of Ω-logical validity.

4.2.1 Axioms[1]

- Empty set:
  $\exists x \forall u(u \notin x)$
- Extensionality:
  $x = y \iff \forall u(u \in x \iff u \in y)$
- Pairing:
  $\exists x \forall u(u \in x \iff u = a \vee u = b)$
- Union:
  $\exists x \forall u[u \in x \iff \exists v(u \in v \wedge v \in a)]$
- Separation:
  $\exists x \forall u[u \in x \iff u \in a \wedge \phi(u)]$
- Power Set:
  $\exists x \forall u(u \in x \iff u \subseteq a)$

[1]For a standard presentation, see Jech (2003). For detailed, historical discussion, see Maddy (1988a).

- Infinity:
  $\exists x \varnothing \in x \wedge \forall u(u \in x \rightarrow \{u\} \in x)$
- Replacement:
  $\forall u \exists! v \psi(u, v) \rightarrow \forall x \exists y (\forall u \in x)(\exists v \in y) \psi(u, v)$
- Choice:
  $\forall u[u \in a \rightarrow \exists v(v \in u)] \wedge \forall u, x[u \in a \wedge x \in a \rightarrow \exists v(v \in u \iff v \in x) \vee \neg v(v \in u \wedge v \in x)] \rightarrow \exists x \forall u[u \in a \rightarrow \exists! v(v \in u \wedge u \in x)]$

4.2.2 *Large Cardinals*

Borel sets of reals are subsets of ω^ω or $\mathbb{R}$, closed under countable intersections and unions.[2] For all ordinals, a, such that $0 < a < \omega_1$, and $b < a$, Σ^0_a denotes the open subsets of ω^ω formed under countable unions of sets in Π^0_b, and Π^0_a denotes the closed subsets of ω^ω formed under countable intersections of Σ^0_b.

Projective sets of reals are subsets of ω^ω, formed by complementations ($\omega^\omega - u$, for $u \subseteq \omega^\omega$) and projections $[p(u) = \{\langle x_1, \ldots, x_n\rangle \in \omega^\omega \mid \exists y \langle x_1, \ldots, x_n, y\rangle \in u\}]$. For all ordinals a, such that $0 < a < \omega$, Π^1_0 denotes closed subsets of ω^ω; Π^1_a is formed by taking complements of the open subsets of ω^ω, Σ^1_a; and Σ^1_{a+1} is formed by taking projections of sets in Π^1_a.

The full power set operation defines the cumulative hierarchy of sets, V, such that $V_0 = \varnothing$; $V_{a+1} = P(V_0)$; and $V_\lambda = \bigcup_{a<\lambda} V_a$.

In the inner model program (cf. Woodin 2010, 2011; Kanamori 2012a,b), the definable power set operation defines the constructible universe, $L(\mathbb{R})$, in the universe of sets V, where the sets are transitive such that $a \in C \iff a \subseteq C$; $L(\mathbb{R}) = V_{\omega+1}$; $L_{a+1}(\mathbb{R}) = \mathrm{Def}(L_a(\mathbb{R}))$; and $L_\lambda(\mathbb{R}) = \bigcup_{a<\lambda}(L_a(\mathbb{R}))$.

Via inner models, Gödel (1940) proves the consistency of the generalized continuum hypothesis, $\aleph_a^{\aleph_a} = \aleph_{a+1}$, as well as the axiom of choice, relative to the axioms of ZFC. However, for a countable transitive set of ordinals, M, in a model of ZF without choice, one can define a generic set, G, such that, for all formulas, ϕ, either ϕ or $\neg\phi$ is forced by a condition, f, in G. Let $M[G] = \bigcup_{a<\kappa} M_a[G]$, such that $M_0[G] = \{G\}$; with $\lambda < \kappa$, $M_\lambda[G] = \bigcup_{a<\lambda} M_a[G]$; and $M_{a+1}[G] = V_a \cap M_a[G]$.[3] G is a Cohen real over M, and comprises a set-forcing extension of M. The relation of set-forcing, $\Vdash$, can then be defined in the ground model, M, such that the forcing condition, f, is a function from a finite subset of ω into {0,1}, and $f \Vdash u \in G$ if $f(u) = 1$ and $f \Vdash u \notin G$ if $f(u) = 0$. The cardinalities of an open dense ground model, M, and a generic extension, G, are identical, only if the countable chain condition (c.c.c.) is satisfied, such that, given a chain – i.e., a linearly ordered subset of a partially ordered (reflexive, antisymmetric, transitive) set – there is a countable,

[2] See Koellner (2013), for the presentation, and for further discussion, of the definitions in this and the subsequent paragraph.

[3] See Kanamori (2012a: 2.1; 2012b: 4.1), for further discussion.

maximal antichain consisting of pairwise incompatible forcing conditions. Via set-forcing extensions, Cohen (1963, 1964) constructs a model of ZF which negates the generalized continuum hypothesis, and thus proves the independence thereof relative to the axioms of ZF.[4]

Gödel (1946/1990: 1–2) proposes that the value of Orey sentences such as the GCH might yet be decidable, if one avails of stronger theories to which new axioms of infinity – i.e., large cardinal axioms – are adjoined.[5] He writes that: 'In set theory, e.g., the successive extensions can be represented by stronger and stronger axioms of infinity. It is certainly impossible to give a combinatorial and decidable characterization of what an axiom of infinity is; but there might exist, e.g., a characterization of the following sort: An axiom of infinity is a proposition which has a certain (decidable) formal structure and which in addition is true. Such a concept of demonstrability might have the required closure property, i.e. the following could be true: Any proof for a set-theoretic theorem in the next higher system above set theory ... is replaceable by a proof from such an axiom of infinity. It is not impossible that for such a concept of demonstrability some completeness theorem would hold which would say that every proposition expressible in set theory is decidable from present axioms plus some true assertion about the largeness of the universe of sets'.

For cardinals, x,a,C, $C \subseteq a$ is closed unbounded in a, if it is closed [if $x < C$ and $\bigcup(C \cap a) = a$, then $a \in C$] and unbounded ($\bigcup C = a$) (Kanamori, op. cit.: 360). A cardinal, S, is stationary in a, if, for any closed unbounded $C \subseteq a$, $C \cap S \neq \emptyset$ (op. cit.). An ideal is a subset of a set closed under countable unions, whereas filters are subsets closed under countable intersections (361). A cardinal κ is regular if the cofinality of κ – comprised of the unions of sets with cardinality less than κ – is identical to κ. Uncountable regular limit cardinals are weakly inaccessible (op. cit.). A strongly inaccessible cardinal is regular and has a strong limit, such that if $\lambda < \kappa$, then $2^{\lambda} < \kappa$ (op. cit.).

Large cardinal axioms are defined by elementary embeddings.[6] Elementary embeddings can be defined thus. For models A,B, and conditions ϕ, j: A $\rightarrow$ B, $\phi\langle a_1, \ldots, a_n\rangle$ in A if and only if $\phi\langle j(a_1), \ldots, j(a_n)\rangle$ in B (363). A measurable cardinal is defined as the ordinal denoted by the critical point of j, crit(j) (Koellner and Woodin 2010: 7). Measurable cardinals are inaccessible (Kanamori, op. cit.).

Let κ be a cardinal, and $\eta > \kappa$ an ordinal. κ is then η-strong, if there is a transitive class M and an elementary embedding, j: V $\rightarrow$ M, such that $\text{crit}(j) = \kappa$, $j(\kappa) > \eta$, and $V_{\eta} \subseteq M$ (Koellner and Woodin, op. cit.).

κ is strong if and only if, for all η, it is η-strong (op. cit.).

[4] See Kanamori (2008), for further discussion.

[5] See Kanamori (2007), for further discussion. Kanamori (op. cit.: 154) notes that Gödel (1931/1986: fn48a) makes a similar appeal to higher-order languages, in his proofs of the incompleteness theorems. The incompleteness theorems are examined in further detail, in Sect. 4.4.2, below.

[6] The definitions in the remainder of this subsection follow the presentations in Koellner and Woodin (2010) and Woodin (2010, 2011).

If A is a class, κ is η-A-strong, if there is a $j : V \rightarrow M$, such that κ is η-strong and $j(A \cap V_\kappa) \cap V_\eta = A \cap V_\eta$ (op. cit.).

κ is a Woodin cardinal, if κ is strongly inaccessible, and for all $A \subseteq V_\kappa$, there is a cardinal $\kappa_A < \kappa$, such that κ_A is η-A-strong, for all η such that κ_η, $\eta < \kappa$ (Koellner and Woodin, op. cit.: 8).

κ is superstrong, if $j : V \rightarrow M$, such that crit(j) $= \kappa$ and $V_{j(\kappa)} \subseteq M$, which entails that there are arbitrarily large Woodin cardinals below κ (op. cit.).

Large cardinal axioms can then be defined as follows.

$\exists x \Phi$ is a large cardinal axiom, because:

(i) Φx is a Σ_2-formula;
(ii) if κ is a cardinal, such that $V \models \Phi(\kappa)$, then κ is strongly inaccessible; and
(iii) for all generic partial orders $\mathbb{P} \in V_\kappa$, $V^{\mathbb{P}} \models \Phi(\kappa)$; I_{NS} is a non-stationary ideal; A^G is the canonical representation of reals in $L(\mathbb{R})$, i.e. the interpretation of A in M[G]; $H(\kappa)$ is comprised of all of the sets whose transitive closure is $< \kappa$ (cf. Rittberg 2015); and $L(\mathbb{R})^{\mathbb{P}max} \models \langle H(\omega_2), \in, I_{NS}, A^G \rangle \models `\phi'$. $\mathbb{P}$ is a homogeneous partial order in $L(\mathbb{R})$, such that the generic extension of $L(\mathbb{R})^{\mathbb{P}}$ inherits the generic invariance, i.e., the absoluteness, of $L(\mathbb{R})$. Thus, $L(\mathbb{R})^{\mathbb{P}max}$ is (i) effectively complete, i.e. invariant under set-forcing extensions; and (ii) maximal, i.e. satisfies all Π_2-sentences and is thus consistent by set-forcing over ground models (Woodin (ms): 28).

Assume ZFC and that there is a proper class of Woodin cardinals; $A \in \mathbb{P}(\mathbb{R}) \cap L(\mathbb{R})$; ϕ is a Π_2-sentence; and V(G), s.t. $\langle H(\omega_2), \in, I_{NS}, A^G \rangle \models `\phi'$: Then, it can be proven that $L(\mathbb{R})^{\mathbb{P}max} \models \langle H(\omega_2), \in, I_{NS}, A^G \rangle \models `\phi'$, where $`\phi' := \exists A \in \Gamma^\infty \langle H(\omega_1), \in, A \rangle \models \psi$.

The axiom of determinacy (AD) states that every set of reals, $a \subseteq \omega^\omega$ is determined, where κ is determined if it is decidable.

Woodin's (1999) Axiom (*) can be thus countenanced:

$AD^{L(\mathbb{R})}$ and $L[(\mathbb{P}\omega_1)]$ is a $\mathbb{P}$max-generic extension of $L(\mathbb{R})$,

from which it can be derived that $2^{\aleph_0} = \aleph_2$. Thus, ¬CH; and so CH is absolutely decidable.

4.2.3 Ω-Logic

For partial orders, $\mathbb{P}$, let $V^{\mathbb{P}} = V^{\mathbb{B}}$, where $\mathbb{B}$ is the regular open completion of ($\mathbb{P}$).[7] $M_a = (V_a)^M$ and $M_a^{\mathbb{B}} = (V_a^{\mathbb{B}})^M = (V_a^{M^{\mathbb{B}}})$. *Sent* denotes a set of sentences in a first-order language of set theory. $T \cup \{\phi\}$ is a set of sentences extending ZFC. *c.t.m* abbreviates the notion of a countable transitive ∈-model. *c.B.a.* abbreviates the notion of a complete Boolean algebra.

[7] The definitions in this section follow the presentation in Bagaria et al. (2006).

Define a $c.B.a.$ in V, such that $V^{\mathbb{B}}$. Let $V_0^{\mathbb{B}} = \emptyset$; $V_\lambda^{\mathbb{B}} = \bigcup_{b<\lambda} V_b^{\mathbb{B}}$, with λ a limit ordinal; $V_{a+1}^{\mathbb{B}} = \{f : X \rightarrow \mathbb{B} \,|X \subseteq V_a^{\mathbb{B}}\}$; and $V^{\mathbb{B}} = \bigcup_{a \in On} V_a^{\mathbb{B}}$.

ϕ is true in $V^{\mathbb{B}}$, if its Boolean-value is $1^{\mathbb{B}}$, if and only if

$V^{\mathbb{B}} \models \phi$ iff $[\![\phi]\!]^{\mathbb{B}} = 1^{\mathbb{B}}$.

Thus, for all ordinals, a, and every $c.B.a.\mathbb{B}$, $V_a^{\mathbb{B}} \equiv (V_a)^{V^{\mathbb{B}}}$ iff for all $x \in V^{\mathbb{B}}$, $\exists y \in V^{\mathbb{B}} [\![x = y]\!]^{\mathbb{B}} = 1^{\mathbb{B}}$ iff $[\![x \in V^{\mathbb{B}}]\!]^{\mathbb{B}} = 1^{\mathbb{B}}$.

Then, $V_a^{\mathbb{B}} \models \phi$ iff $V^{\mathbb{B}} \models$ '$V_a \models \phi$'.

Ω-logical validity can then be defined as follows:

For $T \cup \{\phi\} \subseteq Sent$,

$T \models_\Omega \phi$, if for all ordinals, a, and $c.B.a.\mathbb{B}$, if $V_a^{\mathbb{B}} \models T$, then $V_a^{\mathbb{B}} \models \phi$.

Supposing that there exists a proper class of Woodin cardinals and if $T \cup \{\phi\} \subseteq Sent$, then for all set-forcing conditions, $\mathbb{P}$:

$T \models_\Omega \phi$ iff $V^T \models$ '$T \models_\Omega \phi$',

where $T \models_\Omega \phi \equiv \emptyset \models$ '$T \models_\Omega \phi$'.

The Ω-Conjecture states that $V \models_\Omega \phi$ iff $V^{\mathbb{B}} \models_\Omega \phi$ (Woodin ms). Thus, Ω-logical validity is invariant in all set-forcing extensions of ground models in the set-theoretic multiverse.

The soundness of Ω-Logic is defined by universally Baire sets of reals. For a cardinal, e, let a set A be e-universally Baire, if for all partial orders $\mathbb{P}$ of cardinality e, there exist trees, S and T on ω X λ, such that $A = p[T]$ and if $G \subseteq \mathbb{P}$ is generic, then $p[T]^G = \mathbb{R}^G - p[S]^G$ (Koellner 2013). A is universally Baire, if it is e-universally Baire for all e (op. cit.).

Ω-Logic is sound, such that $V \vdash_\Omega \phi \rightarrow V \models_\Omega \phi$. However, the completeness of Ω-Logic has yet to be resolved.

Finally, in category theory, a category C is comprised of a class Ob(C) of objects a family of arrows for each pair of objects C(A,B) (Venema 2007: 421). A functor from a category C to a category D, **E**: C → D, is an operation mapping objects and arrows of C to objects and arrows of D (422). An endofunctor on C is a functor, **E**: C → C (op. cit.).

A **E**-coalgebra is a pair $\mathbb{A} = (A, \mu)$, with A an object of C referred to as the carrier of $\mathbb{A}$, and μ: A → **E**(A) is an arrow in C, referred to as the transition map of $\mathbb{A}$ (390).

$\mathbb{A} = \langle A, \mu: A \rightarrow \mathbf{E}(A)\rangle$ is dually isomorphic to the category of algebras over the functor μ (417–418). If μ is a functor on categories of sets, then Boolean-algebraic models of Ω-logical validity are isomorphic to coalgebraic models.

The significance of the foregoing is that coalgebraic models may themselves be availed of in order to define modal logic and automata theory. Coalgebras provide therefore a setting in which the Boolean-valued models of set theory, the modal profile of Ω-logical validity, and automata can be interdefined. In what follows, $\mathbb{A}$ will comprise the coalgebraic model – dually isomorphic to the complete Boolean-valued algebras defined in the Ω-Logic of ZFC – in which modal similarity types and automata are definable. As a coalgebraic model of modal logic, $\mathbb{A}$ can be defined as follows (407):

For a set of formulas, Φ, let $\nabla\Phi := \Box \bigvee \Phi \wedge \bigwedge \diamond\Phi$, where $\diamond\Phi$ denotes the set $\{\diamond\phi \mid \phi \in \Phi$ (op. cit.). Then,

$\diamond\phi \equiv \nabla\{\phi, \mathrm{T}\}$,

$\Box\phi \equiv \nabla\varnothing \vee \nabla\phi$ (op. cit.).

Let an **E**-coalgebraic modal model, $\mathbb{A} = \langle \mathrm{S}, \lambda, \mathrm{R}[.]\rangle$, such that $\mathbb{S}$,s $\Vdash \nabla\Phi$ if and only if, for all (some) successors σ of s∈S, $[\Phi, \sigma(\mathrm{s}) \in \mathbf{E}(\Vdash_{\mathbb{A}})]$ (op. cit.).

A coalgebraic model of deterministic automata can be thus defined (391). An automaton is a tuple, $\mathbb{A} = \langle \mathrm{A}, \mathrm{a}_I, \mathrm{C}, \delta, \mathrm{F}\rangle$, such that A is the state space of the automaton $\mathbb{A}$; $\mathrm{a}_I \in$A is the automaton's initial state; C is the coding for the automaton's alphabet, mapping numerals to properties of the natural numbers; δ: A X C → A is a transition function, and F⊆ A is the collection of admissible states, where F maps A to {1,0}, such that F: A → 1 if a∈F and A → 0 if a∉F (op. cit.). The determinacy of coalgebraic automata, the category of which is dually isomorphic to the Set category satisfying Ω-logical consequence, is secured by the existence of Woodin cardinals: Assuming ZFC, that λ is a limit of Woodin cardinals, that there is a generic, set-forcing extension G ⊆ the collapse of $\omega < \lambda$, and that $\mathbb{R}^* = \bigcup\{\mathbb{R}^G[a] \mid \mathrm{a} < \lambda\}$, then $\mathbb{R}^* \models$ the axiom of determinacy (AD) (Koellner and Woodin, op. cit.: 10).

Finally, $\mathbb{A} = \langle \mathrm{A}, \alpha{:}\mathrm{A} \rightarrow \mathbf{E}(\mathrm{A})\rangle$ is dually isomorphic to the category of algebras over the functor α (417–418). For a category C, object A, and endofunctor **E**, define a new arrow, α, s.t. $\alpha{:}\mathbf{E}\mathrm{A} \rightarrow \mathrm{A}$. A homomorphism, f, can further be defined between algebras $\langle \mathrm{A}, \alpha\rangle$, and $\langle \mathrm{B}, \beta\rangle$. Then, for the category of algebras, the following commutative square can be defined: (i) $\mathbf{E}\mathrm{A} \rightarrow \mathbf{E}\mathrm{B}$ ($\mathbf{E}f$); (ii) $\mathbf{E}\mathrm{A} \rightarrow \mathrm{A}$ (α); (iii) $\mathbf{E}\mathrm{B} \rightarrow \mathrm{B}$ (β); and (iv) $\mathrm{A} \rightarrow \mathrm{B}$ (f) (cf. Hughes, 2001: 7–8). The same commutative square holds for the category of coalgebras, such that the latter are defined by inverting the direction of the morphisms in both (ii) [$\mathrm{A} \rightarrow \mathbf{E}\mathrm{A}$ (α)], and (iii) [$\mathrm{B} \rightarrow \mathbf{E}\mathrm{B}$ (β)] (op. cit.).

Thus, $\mathbb{A}$ is the coalgebraic category for modal, deterministic automata, dually isomorphic to the complete Boolean-valued algebraic models of Ω-logical validity, as defined in the category of sets.

4.3 Epistemic Modal Algebras and the Computational Theory of Mind

Beyond the remit of Boolean-valued models of set-theoretic languages, models of epistemic modal algebras are availed of by a number of paradigms in contemporary empirical theorizing, including the computational theory of mind and the theory of quantum computability. In Epistemic Modal Algebra, the topological boolean algebra, A, can be formed by taking the powerset of the topological space, X, defined above; i.e., A = P(X). The domain of A is comprised of formula-terms – eliding propositions with names – assigned to elements of P(X), where the proposition-letters are interpreted as encoding states of information. The top element of the

algebra is denoted '1' and the bottom element is denoted '0'. We interpret modal operators, f(x), – i.e., intensional functions in the algebra – as both concerning topological interiority, as well as reflecting *epistemic possibilities*. An Epistemic Modal-valued Algebraic structure has the form, $F = \langle A, D_{P(X)}, \rho \rangle$, where ρ is a mapping from points in the topological space to elements or regions of the algebraic structure; i.e., $\rho : D_{P(X)} \times D_{P(X)} \rightarrow A$. A model over the Epistemic-Modal Topological Boolean Algebraic structure has the form $M = \langle F, V \rangle$, where $V(a) \leq \rho(a)$ and $V(a,b) \wedge \rho(a, b) \leq V(b)$.[8] For all $x_{x/a,\phi}, y \in A$:

$f(1) = 1;$
$f(x) \leq x;$
$f(x \wedge y) = f(x) \wedge f(y);$
$f[f(x)] = f(x);$
$V(a, a) > 0;$
$V(a, a) = 1;$
$V(a, b) = V(b, a);$
$V(a, b) \wedge V(b, c) \leq V(a, c);$
$V(a = a) = \rho(a, a);$
$V(a, b) \leq f[V(a, b)];$
$V(\neg\phi) = \rho(\neg\phi) - f(\phi);$
$V(\diamond\phi) = \rho\phi - f[-V(\phi)];$
$V(\Box\phi) = f[V(\phi)]$ (cf. Lando, op. cit.).[9]

Marcus (2001) argues that mental representations can be treated as algebraic rules characterizing the computation of operations on variables, where the values of a target domain for the variables are universally quantified over and the function is one-one, mapping a number of inputs to an equivalent number of outputs (35–36). Models of the above algebraic rules can be defined in both classical and weighted, connectionist systems: Both a single and multiple nodes can serve to represent the variables for a target domain (42–45). Temporal synchrony or dynamic variable-bindings are stored in short-term working memory (56–57), while information relevant to long-term variable-bindings are stored in registers (54–56). Examples of the foregoing algebraic rules on variable-binding include both the syntactic concatenation of morphemes and noun phrase reduplication in linguistics (37–39, 70–72), as well as learning algorithms (45–48). Conditions on variable-binding are further examined, including treating the binding relation between variables and values as tensor products – i.e., an application of a multiplicative axiom for variables and their values treated as vectors (53–54, 105–106). In order to account for recursively formed, complex representations, which he refers to as structured propositions, Marcus argues instead that the syntax and semantics of such representations can be modeled via an ordered set of registers, which he refers to as 'treelets' (108).

[8] See Lando (2015), McKinsey (1944) and Rasiowa (1963), for further details.

[9] Note that, in cases of Boolean-valued epistemic topological algebras, models of corresponding coalgebras will be topological (cf. Takeuchi 1985 for further discussion).

A strengthened version of the algebraic rules on variable-binding can be accommodated in models of epistemic modal algebras, when the latter are augmented by cylindrifications, i.e., operators on the algebra simulating the treatment of quantification, and diagonal elements.[10] By contrast to Boolean Algebras with Operators, which are propositional, cylindric algebras define first-order logics. Intuitively, valuation assignments for first-order variables are, in cylindric modal logics, treated as possible worlds of the model, while existential and universal quantifiers are replaced by, respectively, possibility and necessity operators ($\diamond$ and $\Box$) (Venema 2013: 249). For first-order variables, $\{v_i \mid i < \alpha\}$ with α an arbitrary, fixed ordinal, $v_i = v_j$ is replaced by a modal constant $\mathbf{d}_{i,j}$ (op. cit: 250). The following clauses are valid, then, for a model, M, of cylindric modal logic, with $E_{i,j}$ a monadic predicate and T_i for i, j $< \alpha$ a dyadic predicate:

M, w $\Vdash$ p $\iff$ w $\in$ V(p);
M, w $\Vdash \mathbf{d}_{i,j} \iff$ w $\in E_{i,j}$;
M, w $\Vdash \diamond_i \psi \iff$ there is a v with wT_iv and M, v $\vdash \psi$ (252).[11]

Finally, a cylindric modal algebra of dimension α is an algebra, $\mathbb{A} = \langle A, +, \bullet, -, 0, 1, \diamond_i, \mathbf{d}_{ij}\rangle_{i,j<\alpha}$, where $\diamond_i$ is a unary operator which is normal ($\diamond_i 0 = 0$) and additive [$\diamond_i(x + y) = \diamond_i x + \diamond_i y$)] (257).

The philosophical interest of cylindric modal algebras to Marcus' cognitive models of algebraic variable-binding is that variable substitution is treated in the modal algebras as a modal relation, while universal quantification is interpreted as necessitation. The interest of translating universal generalization into operations of epistemic necessitation is, finally, that – by identifying epistemic necessity with apriority – both the algebraic rules for variable-binding and the recursive formation of structured propositions can be seen as operations, the implicit knowledge of which is apriori.

In quantum information theory, let a constructor be a computation defined over physical systems. Constructors entrain nomologically possible transformations from admissible input states to output states (cf. Deutsch 2013). On this approach,

[10] See Henkin et al (op. cit.: 162–163) for the introduction of cylindric algebras, and for the axioms governing the cylindrification operators.

[11] Cylindric frames need further to satisfy the following axioms (op. cit.: 254):

1. $p \rightarrow \diamond_i p$
2. $p \rightarrow \Box_i \diamond_i p$
3. $\diamond_i \diamond_i p \rightarrow \diamond_i p$
4. $\diamond_i \diamond_j p \rightarrow \diamond_j \diamond_i p$
5. $\mathbf{d}_{i,i}$
6. $\diamond_i(\mathbf{d}_{i,j} \wedge p) \rightarrow \Box_i(\mathbf{d}_{i,j} \rightarrow p)$
   [Translating the diagonal element and cylindric (modal) operator into, respectively, monadic and dyadic predicates and universal quantification: $\forall xyz[(T_i xy \wedge E_{i,j} y \wedge T_i xz \wedge E_{i,j} z) \rightarrow y = z]$ (op. cit.)]
7. $\mathbf{d}_{i,j} \iff \diamond_k(\mathbf{d}_{i,k}], \wedge \mathbf{d}_{k,j})$.

information is defined in terms of constructors, i.e., intensional computational properties. The foregoing transformations, as induced by constructors, are referred to as tasks. Because constructors encode the counterfactual to the effect that, were an initial state to be computed over, then the output state would result, modal notions are thus constitutive of the definition of the tasks at issue. There are, further, both topological and algebraic aspects of the foregoing modal approach to quantum computation.[12] The composition of tasks is formed by taking their union, where the union of tasks can be satisfiable while its component tasks might not be. Suppose, e.g., that the information states at issue concern the spin of a particle. A spin-state vector will be the sum of the probabilities that the particle is spinning either upward or downward. Suppose that there are two particles which can be spinning either upward or downward. Both particles can be spinning upward; spinning downward; particle-1 can be spinning upward while particle-2 spins downward; and vice versa. The state vector, V which records the foregoing possibilities – i.e., the superposition of the states – will be equal to the product of the spin-state of particle-1 and the spin-state of particle-2. If the particles are both spinning upward or both spinning downward, then V will be .5. However – relative to the value of each particle vector, referred to as its eigenvalue – the probability that particle-1 will be spinning upward is .5 and the probability that particle-2 will be spinning downward is .5, such that the probability that both will be spinning upward or downward = $.5 \times .5 = .25$. Considered as the superposition of the two states, V will thus be unequal to the product of their eigenvalues, and is said to be entangled. If the indeterminacy evinced by entangled states is interpreted as inconsistency, then the computational properties at issue might further have to be defined on a distribution of epistemic possibilities which permit of hyperintensional distinctions.[13]

4.4 Modal Coalgebraic Automata and the Philosophy of Mathematics

This section examines the philosophical significance of the Boolean-valued models of set-theoretic languages and the modal coalgebraic automata to which they are dually isomorphic. I argue that, similarly to second-order logical consequence, (i) the 'mathematical entanglement' of Ω-logical validity does not undermine its status as a relation of pure logic; and (ii) both the modal profile and model-theoretic characterization of Ω-logical consequence provide a guide to its epistemic

[12]For an examination of the interaction between topos theory and an S4 modal axiomatization of computable functions, see Awodey et al. (2000).

[13]The nature of the indeterminacy in question is examined in Saunders and Wallace (2008), Deutsch (2010), Hawthorne (2010), Wilson (2011), Wallace (2012: 287–289), Lewis (2016: 277–278), and Elohim (ms). For a thorough examination of approaches to the ontology of quantum mechanics, see Arntzenius (2012: ch. 3).

tractability.[14] I argue, then, that there are several considerations adducing in favor of the claim that the interpretation of the concept of set constitutively involves modal notions. The role of the category of modal coalegebraic deterministic automata in (i) characterizing the modal profile of Ω-logical consequence, and (ii) being constitutive of the formal understanding-conditions for the concept of set, provides, then, support for a realist conception of the cumulative hierarchy.

4.4.1 Neo-Logicism

Frege's (1884/1980; 1893/2013) proposal – that cardinal numbers can be explained by specifying an equivalence relation, expressible in the signature of second-order logic and identity, on lower-order representatives for higher-order entities – is the first attempt to provide a foundation for mathematics on the basis of logical axioms rather than rational or empirical intuition. In Frege (1884/1980. cit.: 68) and Wright (1983: 104–105), the number of the concept, **A**, is argued to be identical to the number of the concept, **B**, if and only if there is a one-to-one correspondence between **A** and **B**, i.e., there is a bijective mapping, R, from **A** to **B**. With Nx: a numerical term-forming operator,

- $\forall \mathbf{A} \forall \mathbf{B} \exists R[[Nx: \mathbf{A} = Nx: \mathbf{B} \equiv \exists R[\forall x[\mathbf{A}x \rightarrow \exists y(\mathbf{B}y \wedge Rxy \wedge \forall z(\mathbf{B}z \wedge Rxz \rightarrow y = z))] \wedge \forall y[\mathbf{B}y \rightarrow \exists x(\mathbf{A}x \wedge Rxy \wedge \forall z(\mathbf{A}z \wedge Rzy \rightarrow x = z))]]].$

Frege's Theorem states that the Dedekind-Peano axioms for the language of arithmetic can be derived from the foregoing abstraction principle, as augmented to the signature of second-order logic and identity.[15] Thus, if second-order logic may be counted as pure logic, despite that domains of second-order models are definable via power set operations, then one aspect of the philosophical significance of the abstractionist program consists in its provision of a foundation for classical mathematics on the basis of pure logic as augmented with non-logical implicit definitions expressed by abstraction principles.

There are at least three reasons for which a logic defined in ZFC might not undermine the status of its consequence relation as being logical. The first reason for which the mathematical entanglement of Ω-logical validity might be innocuous is that, as Shapiro (1991: 5.1.4) notes, many mathematical properties cannot be defined within first-order logic, and instead require the expressive resources of second-order logic. For example, the notion of well-foundedness cannot be expressed in a first-order framework, as evinced by considerations of compactness. Let E be a binary relation. Let m be a well-founded model, if there is no infinite sequence, $a_0, \ldots,$

[14] The phrase, 'mathematical entanglement', is owing to Koellner (2010: 2).

[15] Cf. Dedekend (1888/1963) and Peano (1889/1967). See Wright (1983: 154–169) for a proof sketch of Frege's theorem; Boolos (1987) for the formal proof thereof; and Parsons (1964) for an incipient conjecture of the theorem's validity.

a_i, such that $Ea_0, \ldots, Ea_{i+1}$ are all true. If m is well-founded, then there are no infinite-descending E-chains. Suppose that T is a first-order theory containing m, and that, for all natural numbers, n, there is a T with $n + 1$ elements, $a_0, \ldots, a_n$, such that $\langle a_0, a_1\rangle, \ldots, \langle a_n, a_{n-1}\rangle$ are in the extension of E. By compactness, there is an infinite sequence such that $a_0 \ldots a_i$, s.t. $Ea_0, \ldots, Ea_{i+1}$ are all true. So, m is not well-founded.

By contrast, however, well-foundedness can be expressed in a second-order framework:

$\forall X[\exists x Xx \rightarrow \exists x[Xx \wedge \forall y(Xy \rightarrow \neg Eyx)]]$, such that m is well-founded iff every non-empty subset X has an element x, s.t. nothing in X bears E to x.

One aspect of the philosophical significance of well-foundedness is that it provides a distinctively second-order constraint on when the membership relation in a given model is intended. This contrasts with Putnam's (1980) claim, that first-order models *mod* can be intended, if every set s of reals in *mod* is such that an ω-model in *mod* contains s and is constructible, such that – given the Downward Lowenheim-Skolem theorem[16] – if *mod* is non-constructible but has a submodel satisfying 's is constructible', then the model is non-well-founded and yet must be intended. The claim depends on the assumption that general understanding-conditions and conditions on intendedness must be co-extensive, to which I will return in Sect. 4.4.2

A second reason for which Ω-logic's mathematical entanglement might not be pernicious, such that the consequence relation specified in the Ω-logic might be genuinely logical, may again be appreciated by its comparison with second-order logic. Shapiro (1998) defines the model-theoretic characterization of logical consequence as follows:

'(10) Φ is a logical consequence of [a model] Γ if Φ holds in all possibilities under every interpretation of the nonlogical terminology which holds in Γ' (148).

A condition on the foregoing is referred to as the 'isomorphism property', according to which 'if two models M, M' are isomorphic vis-a-vis the nonlogical items in a formula Φ, then M satisfies Φ if and only if M' satisfies Φ' (151).

Shapiro argues, then, that the consequence relation specified using second-order resources is logical, because of its modal and epistemic profiles. The epistemic tractability of second-order validity consists in 'typical soundness theorems, where one shows that a given deductive system is 'truth-preserving' (154). He writes that: '[I]f we know that a model is a good mathematical model of logical consequence (10), then we know that we won't go wrong using a sound deductive system. Also, we can know that an argument is a logical consequence . . . via a set-theoretic proof in the metatheory' (154–155).

The modal profile of second-order validity provides a second means of accounting for the property's epistemic tractability. Shapiro argues, e.g., that: 'If the isomorphism property holds, then in evaluating sentences and arguments, the only 'possibility' we need to 'vary' is the size of the universe. If enough sizes are

[16]For any first-order model M, M has a submodel M' whose domain is at most denumerably infinite, s.t. for all assignments s on, and formulas $\phi(x)$ in, M', $M,s \Vdash \phi(x) \iff M',s \Vdash \phi(x)$.

represented in the universe of models, then the modal nature of logical consequence will be registered … [T]he only 'modality' we keep is 'possible size', which is relegated to the set-theoretic metatheory' (152).

Shapiro's remarks about the considerations adducing in favor of the logicality of non-effective, second-order validity generalize to Ω-logical validity. In the previous section, the modal profile of Ω-logical validity was codified by the dual isomorphism between complete Boolean-valued algebraic models of Ω-logic and the category, $\mathbb{A}$, of coalgebraic modal logics. As with Shapiro's definition of logical consequence, where Φ holds in all possibilities in the universe of models and the possibilities concern the 'possible size' in the set-theoretic metatheory, the Ω-Conjecture states that $V \models_\Omega \phi$ iff $V^{\mathbb{B}} \models_\Omega \phi$, such that Ω-logical validity is invariant in all set-forcing extensions of ground models in the set-theoretic multiverse.

Finally, the epistemic tractability of Ω-logical validity is secured, both – as on Shapiro's account of second-order logical consequence – by its soundness, but also by its isomorphism to the coalgebraic category of deterministic automata, where the determinacy thereof is again secured by the existence of Woodin cardinals.

4.4.2 Set-Theoretic Realism

In this section, I argue, finally, that the modal profile of Ω-logic can be availed of in order to account for the understanding-conditions of the concept of set, and thus crucially serve as part of the argument for set-theoretic realism.

Putnam (op. cit.: 473–474) argues that defining models of first-order theories is sufficient for both understanding and specifying an intended interpretation of the latter. Wright (1985: 124–125) argues, by contrast, that understanding-conditions for mathematical concepts cannot be exhausted by the axioms for the theories thereof, even on the intended interpretations of the theories. He suggests, e.g., that:

'[I]f there really were uncountable sets, their existence would surely have to flow from the concept of set, as intuitively satisfactorily explained. Here, there is, as it seems to me, no assumption that the content of the ZF-axioms cannot exceed what is invariant under all their classical models. [Benacerraf] writes, e.g., that: 'It is granted that they are to have their 'intended interpretation': 'e' is to mean set-membership. Even so, and conceived as encoding the intuitive concept of set, they fail to entail the existence of uncountable sets. So how can it be true that there are such sets? Benacerraf's reply is that the ZF-axioms are indeed faithful to the relevant informal notions only if, in addition to ensuring that 'E' means set-membership, we interpret them so as to observe the constraint that 'the universal quantifier has to mean all or at least all sets' (p. 103). It follows, of course, that if the concept of set does determine a background against which Cantor's theorem, under its intended interpretation, is sound, there is more to the concept of set that can be explained by communication of the intended sense of 'e' and the stipulation that the ZF-axioms are to hold. And the residue is contained, presumably, in the informal explanations to which, Benacerraf reminds us, Zermelo intended his formalization to answer. At least, this must be so if

the 'intuitive concept of set' is capable of being explained at all. Yet it is notable that Benacerraf nowhere ventures to supply the missing informal explanation – the story which will pack enough into the extension of 'all sets' to yield Cantor's theorem, under its intended interpretation, as a highly non-trivial corollary'(op. cit).

In order to provide the foregoing explanation in virtue of which the concept of set can be shown to be associated with a realistic notion of the cumulative hierarchy, I will argue that there are several points in the model theory and epistemology of set-theoretic languages at which the interpretation of the concept of set constitutively involves modal notions. The aim of the section will thus be to provide a modal foundation for mathematical platonism.

One point is in the coding of the signature of the theory, T, in which Gödel's incompleteness theorems are proved (cf. Halbach and Visser 2014). Relative to,

(i) a choice of coding for an ω-complete, recursively axiomatizable language, L, of T – i.e. a mapping between properties of numbers and properties of terms and formulas in L;
(ii) a predicate, phi; and
(iii) a fixed-point construction:

Let phi express the property of 'being provable', and define (iii) such that, for all consistent theories T of L, there are sentences, p_{phi}, corresponding to each formula, phi(x), in T, s.t. for 'm' := p_{phi},

$\vdash_T p_{phi}$ iff phi(m).

One can then construct a sentence, 'm' := ¬phi(m), such that L is incomplete (the first incompleteness theorem).

Moreover, L cannot prove its own consistency:

If:

$\vdash_T$ 'm' iff ¬phi(m),

Then:

$\vdash_T C \rightarrow m$.

Thus, L is consistent only if L is inconsistent (the second incompleteness theorem).

In the foregoing, the choice of coding bridges the numerals in the language with the properties of the target numbers. The choice of coding is therefore intensional, and has been marshalled in order to argue that the very notion of syntactic computability – via the equivalence class of partial recursive functions, λ-definable terms, and the transition functions of discrete-state automata such as Turing machines – is constitutively semantic (cf. Rescorla 2015). Further points at which intensionality can be witnessed in the phenomenon of self-reference in arithmetic are introduced by Reinhardt (1986). Reinhardt (op. cit.: 470–472) argues that the provability predicate can be defined relative to the minds of particular agents – similarly to Quine's (1968) and Lewis' (1979) suggestion that possible worlds can be centered by defining them relative to parameters ranging over tuples of spacetime coordinates or agents and locations – and that a theoretical identity statement can be established for the concept of the foregoing minds and the concept of a computable system.

In the previous section, intensional computational properties were defined via modal coalgebraic deterministic automata, where the coalgebraic categories are dually isomorphic to the category of sets in which Ω-logical validity was defined. Coalgebraic modal logic was shown to elucidate the modal profile of Ω-logical consequence in the Boolean-valued algebraic models of set theory. The intensionality witnessed by the choice of coding may therefore be further witnessed by the modal automata specified in the foregoing coalgebraic logic.

A second point at which understanding-conditions may be shown to be constitutively modal can be witnessed by the conditions on the epistemic entitlement to assume that the language in which Gödel's second incompleteness theorem is proved is consistent (cf. Dummett 1963/1978; Wright 1985). Wright (op. cit.: 91, fn.9) suggests that '[T]o treat [a] proof as establishing consistency is implicitly to exclude any doubt ... about the consistency of first-order number theory'. Wright's elaboration of the notion of epistemic entitlement, appeals to a notion of rational 'trust', which he argues is recorded by the calculation of 'expected epistemic utility' in the setting of decision theory (2004; 2014: 226, 241). Wright notes that the rational trust subserving epistemic entitlement will be pragmatic, and makes the intriguing point that 'pragmatic reasons are not a special genre of reason, to be contrasted with e.g. epistemic, prudential, and moral reasons' (2012: 484). Crucially, however, the very idea of expected epistemic utility in the setting of decision theory makes implicit appeal to the notion of possible worlds, where the latter can again be determined by the coalgebraic logic for modal automata.

A third consideration adducing in favor of the thought that grasp of the concept of set might constitutively possess a modal profile is that the concept can be defined as an intension – i.e., a function from possible worlds to extensions. The modal similarity types in the coalgebraic modal logic may then be interpreted as dynamic-interpretational modalities, where the dynamic-interpretational modal operator has been argued to entrain the possible reinterpretations both of the domains of the theory's quantifiers (cf. Fine 2005, 2006), as well as of the intensions of non-logical concepts, such as the membership relation (cf. Uzquiano 2015).[17]

The fourth consideration avails directly of the modal profile of Ω-logical consequence. While the above dynamic-interpretational modality will suffice for

[17]For an examination of the philosophical significance of modal coalgebraic automata beyond the philosophy of mathematics, see Baltag (2003). Baltag (op. cit.) proffers a colagebraic semantics for dynamic-epistemic logic, where coalgebraic functors are intended to record the informational dynamics of single- and multi-agent systems. For an algebraic characterization of dynamic-epistemic logic, see Kurz and Palmigiano (2013). For further discussion, see Elohim (ms). The latter proceeds by examining undecidable sentences via the epistemic interpretation of multi-dimensional intensional semantics. See Reinhardt (1974), for a similar epistemic interpretation of set-theoretic languages, in order to examine the reduction of the incompleteness of undecidable sentences on the counterfactual supposition that the language is augmented by stronger axioms of infinity; and Maddy (1988b), for critical discussion. Chihara (2004) argues, as well, that conceptual possibilities can be treated as imaginary situations with regard to the construction of open-sentence tokens, where the latter can then be availed of in order to define nominalistically adequate arithmetic properties.

possible reinterpretations of mathematical terms, the absoluteness and generic invariance of the consequence relation is such that, if the Ω-conjecture is true, then Ω-logical validity is invariant in all possible set-forcing extensions of ground models in the set-theoretic multiverse. The truth of the Ω-conjecture would thereby place an indefeasible necessary condition on a formal understanding of the intension for the concept of set.

4.5 Concluding Remarks

In this essay I have examined the philosophical significance of the isomorphism between Boolean-valued algebraic models of modal Ω-logic and modal coalgebraic models of automata. I argued that – as with the property of validity in second-order logic – Ω-logical validity is genuinely logical, and thus entails a type of neo-logicism in the foundations of mathematics. I argued, then, that modal coalegebraic deterministic automata, which characterize the modal profile of Ω-logical consequence, are constitutive of the interpretation of mathematical concepts such as the membership relation. The philosophical significance of modal Ω-logic is thus that it can be availed of to vindicate both a neo-logicist foundation for set theory and a realist interpretation of the cumulative hierarchy of sets.

References

Arntzenius, F. 2012. *Space, Time, and Stuff*. Oxford: Oxford University Press.

Awodey, S., L. Birkedal, and D. Scott. 2000. Local Realizability Toposes and a Modal Logic for Computability. Technical Report No. CMU-PHIL-99.

Bagaria, J., N. Castells, and P. Larson. 2006. An Ω-logic Primer. *Trends in Mathematics: Set Theory*. Basel: Birkhäuser Verlag.

Baltag, A. 2003. A coalgebraic semantics for epistemic programs. *Electronic Notes in Theoretical Computer Science* 82: 1.

Boolos, G. 1987. The Consistency of Frege's *Foundations of Arithmetic*. In *On Being and Saying*, ed. J.J. Thomson. Cambridge: MIT Press.

Chihara, C. 2004. *A Structural Account of Mathematics*. Oxford: Oxford University Press.

Cohen, P.J. 1963. The independence of the continuum hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 50 (6): 1143–1148.

Cohen, P.J. 1964. The independence of the continuum hypothesis, II. *Proceedings of the National Academy of Sciences of the United States of America* 51 (1): 105–110.

Dedekend, R. 1888/1963. Was sind und was sollen die Zahlen? In *Essays on the Theory of Numbers*. Trans. and ed. W. Beman. New York: Dover.

Deutsch, D. 2010. Apart from Universes. In *Many Worlds? Everett, Quantum Theory, and Reality*, ed. S. Saunders, J. Barrett, A. Kent, and D. Wallace. Oxford: Oxford University Press.

Deutsch, D. 2013. Constructor theory. *Synthese* 190: 4331–4359.

Dummett, M. 1963/1978. The Philosophical Significance of Gödel's Theorem. In *Truth and Other Enigmas*, ed. M. Dummett. Cambridge: Harvard University Press.

Fine, K. 2005. Our Knowledge of Mathematical Objects. In *Oxford Studies in Epistemology*, vol. 1, ed. T. Gendler and J. Hawthorne. Oxford: Oxford University Press.

Fine, K. 2006. Relatively Unrestricted Quantification. In *Absolute Generality*, ed. A. Rayo and G. Uzquiano. Oxford: Oxford University Press.
Frege, G. 1884/1980. *The Foundations of Arithmetic*, 2nd ed. Trans. J.L. Austin. Evanston: Northwestern University Press.
Frege, G. 1893/2013. *Basic Laws of Arithmetic,* vol. I–II. Trans. and ed. P. Ebert, M. Rossberg, C. Wright, and R. Cook. Oxford: Oxford University Press, Evanston, Illinois.
Gödel, K. 1931/1986. On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I. In *Collected Works*, vol. I, ed. S. Feferman, J. Dawson, S. Kleene, G. Moore, R. Solovay, and J. van Heijenoort. Oxford: Oxford University Press.
Gödel, K. 1940. *The consistency of the axiom of choice and of the generalized continuum hypothesis with the axioms of set theory*. Princeton: Princeton University Press.
Gödel, K. 1946/1990. Remarks before the Princeton Bicentennial Conference on Problems in Mathematics. In *Collected Works*, vol. II, ed. S. Feferman, J. Dawson, S. Kleene, G. Moore, R. Solovay, and J. van Heijenoort. Oxford: Oxford University Press.
Halbach, V., and A. Visser. 2014. Self-reference in arithmetic I. *Review of Symbolic Logic* 7: 4.
Hawthorne, J. 2010. A Metaphysician Looks at the Everett Interpretation. In *Many Worlds? Everett, Quantum Theory, and Reality*, ed. S. Saunders, J. Barrett, A. Kent, and D. Wallace. Oxford: Oxford University Press.
Henkin, L., J.D. Monk, and A. Tarski. 1971. *Cylindric Algebras*, Part I. Amsterdam: North-Holland.
Hughes, J. 2001. A study of categories of algebras and Coalgebras. PhD thesis, Department of Philosophy, Carnegie Mellon University, Pittsburgh, May, 2001.
Jech, T. 2003. *Set Theory*, 3rd Millennium ed. Berlin/Heidelberg: Springer.
Kanamori, A. 2007. Gödel and set theory. *Bulletin of Symbolic Logic* 13: 2.
Kanamori, A. 2008. Cohen and set theory. *Bulletin of Symbolic Logic* 14: 3.
Kanamori, A. 2012a. Large Cardinals with Forcing. In *Handbook of the History of Logic: Sets and Extensions in the Twentieth Century*, ed. D. Gabbay, A. Kanamori, and J. Woods. Amsterdam: Elsevier.
Kanamori, A. 2012b. Set theory from Cantor to Cohen. In *Handbook of the History of Logic: Sets and Extensions in the Twentieth Century*, ed. D. Gabbay, A. Kanamori, and J. Woods. Amsterdam: Elsevier.
Koellner, P. 2010. On strong logics of first and second order. *Bulletin of Symbolic Logic* 16: 1.
Koellner, P. 2013. Large cardinals and determinacy. In *Stanford encyclopedia of philosophy*, (Winter 2013 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/entries/large-cardinals-determinacy/
Koellner, P., and W.H. Woodin. 2010. Large Cardinals from Determinacy. In *Handbook of Set Theory*, vol. 3, ed. M. Foreman and A. Kanamori. Dordrecht/Heidelberg: Springer.
Kurz, A., and A. Palmigiano. 2013. Epistemic updates on algebras. *Logical Methods in Computer Science* 9(4): 17.
Lando, T. 2015. First order S4 and its measure-theoretic semantics. *Annals of Pure and Applied Logic* 166: 187–218.
Lewis, D. 1979. Attitudes De Dicto and De Se. *Philosophical Review* 88: 4.
Lewis, P. 2016. *Quantum Ontology*. New York: Oxford University Press.
Maddy, P. 1988a. Believing the axioms I. *Journal of Symbolic Logic* 53: 2.
Maddy, P. 1988b. Believing the axioms II. *Journal of Symbolic Logic* 53: 3.
Marcus, G. 2001. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge: MIT Press.
McKinsey, J., and A. Tarski. 1944. The algebra of topology. *The Annals of Mathematics, Second Series* 45: 1.
Parsons, Ch. 1964. Frege's theory of number. In *Philosophy in America*, ed. Max Black. London: Allen and Unwin.
Peano, G. 1889/1967. The principles of arithmetic, Presented by a New Method (Trans. J. van Heijenoort). In J. van Heijenoort ed. 1967., 'From Frege to Gödel'. Cambridge: Harvard University Press.

Putnam, H. 1980. Models and reality. *Journal of Symbolic Logic* 45: 3.
Quine, W.V. 1968. Propositional objects. *Crítica* 2: 5.
Rasiowa, H. 1963. On modal theories. *Acta Philosophica Fennica* 16: 123–136.
Reinhardt, W. 1974. Remarks on Reflection Principles, Large Cardinals, and Elementary Embeddings. In *Proceedings of Symposia in Pure Mathematics, Vol. 13, Part 2: Axiomatic Set Theory*, ed. T. Jech. Providence: American Mathematical Society.
Reinhardt, W. 1986. Epistemic theories and the interpretation of Gödel's incompleteness theorems. *Journal of Philosophical Logic* 15: 4.
Rescorla, M. 2015. The representational foundations of computation. *Philosophia Mathematica*. https://doi.org/10.1093/philmat/nkv009
Rittberg, C. 2015. How woodin changed his mind: new thoughts on the continuum hypothesis. *Archive for History of Exact Sciences* 69: 2.
Saunders, S., and D. Wallace. 2008. Branching and uncertainty. *British Journal for the Philosophy of Science* 59: 293–305.
Shapiro, S. 1991. *Foundations Without Foundationalism*. Oxford: Oxford University Press.
Shapiro, S. 1998. Logical Consequence: Models and Modality. In *The Philosophy of Mathematics Today*, ed. M. Schirn. Oxford: Oxford University Press.
Takeuchi, M. 1985. Topological coalgebras. *Journal of Algebra* 97: 505–539.
Uzquiano, G. 2015. Varieties of indefinite extensibility. *Notre Dame Journal of Formal Logic* 58: 1.
Venema, Y. 2007. Algebras and coalgebras. In *Handbook of Modal Logic*, ed. P. Blackburn, J. van Benthem, and F. Wolter. Amsterdam: Elsevier.
Venema, Y. 2013. Cylindric Modal Logic. In *Cylindric-Like Algebras and Algebraic Logic*, ed. H. Andréka, M. Ferenczi, and I. Németi. Berlin/Heidelberg: János Bolyai Mathematical Society/Springer.
Wallace, D. 2012. *The Emergent Multiverse*. Oxford: Oxford University Press.
Wilson, A. 2011. Macroscopic ontology in everettian quantum mechanics. *Philosophical Quarterly* 61: 243.
Woodin, W.H. 1999. *The Axiom of Determinacy, Forcing Axioms, and the Nonstationary Ideal*. Berlin/New York, de Gruyter.
Woodin, W.H. 2010. Strong Axioms of Infinity and the Search for V. In *Proceedings of the International Congress of Mathematicians*. Hyderabad, India.
Woodin, W.H. 2011. The Realm of the Infinite. In *Infinity: New Research Frontiers*, ed. M. Heller and W.H. Woodin. Cambridge: Cambridge University Press.
Woodin, W.H. ms. The Ω Conjecture.
Wright, C. 1983. *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.
Wright, C. 1985. Skolem and the sceptic. *Proceedings of the Aristotelian Society, Supplementary Volume* 59: 85–138.
Wright, C. 2004. Warrant for nothing (and foundations for free)? *Proceedings of the Aristotelian Society, Supplementary Volume* 78: 1.
Wright, C. 2012. Replies, Part IV: Warrant Transmission and Entitlement. In *Mind, Meaning and Knowledge*, ed. A. Coliva. Oxford: Oxford University Press.
Wright, C. 2014. On Epistemic Entitlement II. In *Scepticism and Perceptual Justification*, ed. D. Dodd and E. Zardini. New York: Oxford University Press.

Chapter 5
What Arrow's Information Paradox Says (to Philosophers)

Mario Piazza and Marco Pedicini

Abstract Arrow's information paradox features the most radical kind of information asymmetry by diagnosing an inherent conflict between two parties inclined to exchange information. In this paper, we argue that this paradox is more richly textured than generally supposed by current economic discussion on it and that its meaning encroaches on philosophy. In particular, we uncovers the 'epistemic' and more genuine version of the paradox, which looms on our cognitive lives like a sort of tax on curiosity. Finally, we sketch the relation between Arrow's information paradox and the notion of zero-knowledge proofs in cryptography: roughly speaking, zero-knowledge proofs are protocols that enable a prover to convince a verifier that a statement is true, without conveying any additional information.

Keywords Information asymmetry · Arrow's information paradox · Zero-knowledge proofs · Meno's paradox of inquiry · Shannon's communcication model

5.1 Preamble

In the economic literature, the term 'information asymmetry' refers to any condition in which one party in a transaction has more or better information than the other Akerloff (1970). This asymmetry in contracting is a familiar, ubiquitous and inescapable phenomenon in everyday economic life. As Joseph Stiglitz describes it:

M. Piazza (✉)
Classe di Lettere e Filosofia, Scuola Normale Superiore, Pisa, Italia
e-mail: mario.piazza@sns.it

M. Pedicini
Department of Mathematics and Physics, Roma Tre University, Roma, Italy
e-mail: marco.pedicini@uniroma3.it

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_5

> [...] the person buying insurance knows more about his health [...]; the owner of a car knows more about the car than potential buyers; the owner of a firm knows more about the firm that a potential investor; the borrower knows more about his risk and risk taking than the lender (Stiglitz 2002, p. 465).

The most radical type of information asymmetry emerges when a transaction concerns information itself. This special type of asymmetry was brought to the fore by Kenneth Arrow in his 1962 seminal article "Economic Welfare and the Allocation of Resources for Invention". He noticed that contracting over information generates a general puzzle:

> There is a fundamental paradox in the determination of demand for information; its value for the purchaser is not known until he has the information, but then he has in effect acquired it without cost (Arrow 1962, p. 615)

The situation Arrow has in mind is this: one potential buyer (B) wants to buy the information X from one seller (S), the information holder. However, since B *ignores* (the content of) X, she also ignores whether X is worth its price. Thus, S is reluctant to disclose X to B *prior* to her purchase. Indeed, S is afraid of not being compensated by B, under the psychological assumption that the willingness to pay for information drops radically after one is told what the information is about (and under the trivial assumption that once information is given, it cannot be withdrawn).[1] This is Arrow's information paradox (**AIP**, henceforth).

The contours of Arrow's general view are as follows: (1) information itself is a commodity, although a *sui generis* one, being intangible and satisfying peculiar properties with respect to the optimality of its allocation; (2) in a free-market economy the lack of protection for invention leads to the lack of innovative effort, namely the goal of inventing things is to generate intellectual property rights. Therefore, (3) "precisely to the extent that it is successful there is an *underutilisation* of information" (Arrow 1962, p. 617, our emphasis).

Current discussion on **AIP** has a practical orientation at the expense of a host of important theoretical issues. The conceptualisation of the paradox has been indeed confined to the economics of information literature, where **AIP** is typically interpreted as an argument in favour of an unconditional demand for patent protection and exclusive rights over information, through centralised institutions (Merges 1994; Anton and Yao 2002; Thambisetty 2007).[2] Needless to say, Arrow's thought that the whole process of invention produces as *desirable* effect the "underutilisation of information" was a revolutionary one, contravening the intuition that this underutilisation, as a such, does not give us much to celebrate. Clearly there is a morass of economic and legal issues here, issues to which is impossible to do justice in this paper. Rather, our concern is to argue that **AIP** should not be left only to the economist's perspective. The important point for our purposes is that

[1] Arrow also assumes that information is *indivisible*, i.e. it cannot conveyed in parts that constitute evidence that the information has value.

[2] As Merges puts it: "Arrow has pointed out in his "paradox of information" without a property right, the licensor is in a pickle" (Merges 1994, p. 2657).

such a perspective has dismissed some of the most pressing *epistemic* questions associated with **AIP**, so that the bearing of the paradox has been masked. In other word, we claim that the real value of **AIP** is more epistemic than pragmatic and that its center of gravity should be detected on the buyer's side. Not despite but because the paradox is unsophisticated, it involves something crucial and general about the nature of information itself.

This is the roadmap. In Sect. 5.2, we deal with **AIP** from a game-theoretical point of view, by making assumptions only about agent's psychology and knowledge. In particular, we introduce a distinction between *two* versions of the paradox, spelling out a natural bifurcation of the behaviour of the agent who seeks out information. Then, in Sect. 5.3, we consider the paradox in a classical logic setting which forces a particular reading of it. In Sect. 5.4, we will show that one version of **AIP** is isomorphic to the Paradox of Inquiry in Plato's *Meno*, while not sharing the moral that may be drawn. In Sect. 5.5, our aim is to embed **AIP** in Shannon's model of communication, which provides us a means of expressing intellectual property rights as *noise*. In Sect. 5.6, we turn to some exploratory remarks about how a solution of Arrow's paradox shapes up via the notion of zero-knowledge proof. Then, in Sect. 5.7, we very briefly take stock.

5.2 The Hobbesian and the Epistemic View of AIP

Let us start by unfolding the paradox. A look at the economic literature shows that the mainstream understanding of **AIP** amounts to what we might call the "Hobbesian view" (**H-AIP**). Crucially, what **H-AIP** presupposes is that the potential buyer B is a *dishonest* agent in that B is not willing to pay for X after getting it (as Arrow writes, B acquires the information 'without cost'). Then, the paradox is triggered by a unregimented tension which takes on pungent practical relevance: on one side stands B who wants to know the information X available to S; on the other side stands S who wants to keep this information *secret* from B until the end of transaction. Thus, eventually, either B still ignores X, or B is guilty of cheating, inasmuch as B does not pay for X once declassified. In sum, **H-AIP** features **AIP** as something which, at bottom, concerns the risk or the fear of being cheated in transferring information.

Arguably, the monopoly of this reading of **AIP** explains why the discussion of it has been so far restricted to the economics of information literature, which fosters the idea that the information holder has good reasons to be on the defensive under the threat of misappropriation of information. Important thought it is, however, the risk or the fear of being cheated in transferring information and the consequent precautions cannot be the whole story about the paradox. In short, we submit that the Hobbesian reading is not the only game in town. For philosophers the stakes should be different, since they may find other reasons to be interested in the fate of the paradox: what is properly at issue is not the price of dishonesty, but that of honesty.

Hence let us assume that in Arrow's scenario the buyer B is an *honest* agent, namely that B compensates S for X *anyway*. To prepare ourselves for the philosophical issue to come, we need now to introduce a kind of taxonomy for honesty, a taxonomy which in itself is not philosophically charged. It is simply expressed in terms of the temporality of the action of buying. If B is honest, then one of the following three scenarios may happen:

(1) B *ex ante* buys X, i.e., B first pays for the information X, and then she gets X (i.e., B knows X *after* the closing of the transaction);
(2) B *ex post* buys X, i.e., B has the information X straight away and after she pays for X. (i.e., B knows X *before* the closing of the transaction);
(3) B *ex synchro* buys X, i.e., B gets the information X simultaneously with the payment (i.e., B knows X *simultaneously* to the closing of the transaction).

Please, take note of two obvious points. To say that the buyer *ex ante* (or *ex synchro*) buys the information X is to say nothing more that she buys X *blindly* (i.e., under uncertainty), and to say that the buyer *ex post* buys X is to say that she buys X *without needing* it (B *already* knows X).

Our claim is that the real dimension of **AIP** is *epistemic*, and so the paradox properly falls within the boundaries of epistemology. 'Epistemic view' is the name we give to the following version of the paradox:

(**E-AIP**) Assume that B is honest. Then, B's purchase of X is either *blind* (being *ex ante* or *ex synchro*), or *unnecessary* (being *ex post*).

It is worth observing that **E-AIP** easily generalises. The action of buying the information X, indeed, can be seen as one particular case of the general case of 'performing an action A to get X': is the action A worth its effort? **E-AIP** tells us that this effort is blind *before* knowing X, while it is unnecessary *after* knowing X.

5.3 AIP and Classical Logic

We can see that under classical (and intuitionistic) logic the option *ex synchro* vanishes. Let us suppose that B and S underwrite such an agreement:

B: 'I will pay for X, provided you disclose it';
S: 'I will disclose X, provided you pay for it'.

One may say that the agreement between them is well-balanced, inasmuch as its expected outcome is that B acquires X blindly but not in advance, whereas S gets a compensation for X without suspense, so to say.

Now, let β and α two atomic sentences standing for 'S discloses the information X' and 'B pays for X', respectively. B's commitment is represented as the implication $\beta \to \alpha$ and S's commitment is represented as the converse $\alpha \to \beta$. Yet, this agreement is not congenial to classical propositional logic (and intuitionistic logic as well) because the commitments of the two agents do not suffice to deduce

the conjunction $\alpha \wedge \beta$:

$$(\beta \rightarrow \alpha) \wedge (\alpha \rightarrow \beta) \nvdash \alpha \wedge \beta$$

In point of fact, the classical (and intuitionistic) *modus ponens*: $\alpha \rightarrow \beta, \alpha \vdash \beta$ expresses the fact that S discloses X, but only *after* B has paid for X, whereas the *modus ponens*: $\beta \rightarrow \alpha, \beta \vdash \alpha$ means that B pays for X, but only *after* S has disclosed X. In other terms, under classical and intuitionistic logic, one of the two parties must make the first move in order for the agreement to be effective:

(1) if B makes the first move, then B *ex ante* pays for X;
(2) if S makes the first move, then B *ex post* pays for X.

5.4 Platonizing AIP

Plato's readers will hear a Platonic echo. **E-AIP** is isomorphic to the Paradox of Inquiry in *Meno*, a paradox which is like a Zenonian argument against motion. At a certain point in the *Meno*, Plato makes Socrates say that:

> it is impossible for a person to search either for what he knows or for what he doesn't know: he cannot search for what he knows, since he knows it and that makes the search unnecessary, and he can't search for what he doesn't know either, since he doesn't even know what it is he's going to search for (*Meno*, 80e2–5).

Keeping the very same structure of the Paradox of Inquiry, **E-AIP** runs as follows:

> it is impossible for a person to buy either the information he knows or the information he doesn't know: he cannot buy the information he knows, since he knows it and that makes buying it unnecessary, and he can't buy the information he does not know either, since he doesn't even know what it is he's going to buy.

The Paradox of Inquiry relies heavily on a claim about knowledge that is simply false, namely that knowledge is an *all-or-nothing* affair. In this sense, the paradox works as a *reductio ad absurdum*: *if* knowledge is an all-or-nothing affair, *then* inquiry is impossible. Plato's solution is to admit that one can grasp something *partially*: true beliefs have a status intermediate between blank ignorance and full knowledge. Hence, one could have the temptation to make a similar diagnosis for **E-AIP** by saying that the problem rests on the contentious claim that information is an all-or-nothing thing: *if* information is an all-or-nothing thing, *then* buying it is impossible. In sum, one might think to take comfort in the notion of *partial information*: the buyer can access incomplete or partial information X^{-} before buying the complete one X.

At this point, it would be nice to have an account of the slippery notion of partial information. But apart from the basic difficulty in characterizating it, the trouble behind the appeal to partiality is easy enough to state. Assume B wants to buy X

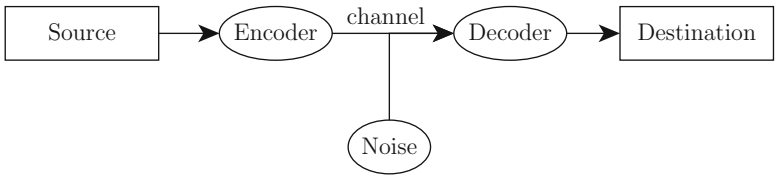

Fig. 5.1 Shannon's communication model

from S and in the meanwhile receives from S some partial information X^-. Agreed, B is told about X^-. But she wants to know X, and X by definition does not coincide with X^-. So, we end up with the fact that either B *ex ante* buys X,[3] or B *ex post* buys X. If B *ex ante* buys X, then B buys X *still* blindly; if B *ex post* buys X, then:

(a) X^- is payoff-irrelevant for S;
(b) the purchase of X is still *unnecessary*.

The conclusion is that one cannot escape the dilemma triggered by **E-AIP** by unpacking information into partial information.

5.5 Shannon Meets Arrow

So far our analysis of **AIP** has only incorporated assumptions about agent's psychology and knowledge, without telling anything about the resources of computation and communication available to the agent. In the original description of **AIP**, information is transmitted *instantly*: **AIP** does not trot out the process whereby the information moves via a channel from one agent to another. Nor **AIP** takes into account some dysfunctional factor in the very transmission. To our knowledge, in the literature there is no description of **AIP** in terms of Shannon's Communication Model (**SCM**) (Shannon 1948).

It is well-known that, according to **SCM**, communication is a transfer process between an information source and a destination. Diagrammatically, information source and destination are at the opposite ends of a chain, Fig. 5.1. The source creates the message which travels along some physical medium – the channel – until it reaches its destination. However, noise can affect almost anywhere the communication process as an unwelcome addition to the message. As dysfunctional factor, noise may prevent the message from reaching its destination or may lead to the message received being different from that sent. To reduce the effect of

[3]This situation is realised by the purchase of the content of online newspapers with paywall systems: these display an article title and a few paragraphs (X^-) before prompting the reader to pay for X.

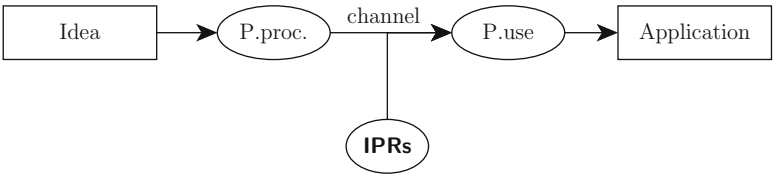

Fig. 5.2 Parallel between **SCM** and the **IPRs** assessment

noise a kind of symmetry is required: an *encoder* is placed between the source and the channel, while a *decoder* is placed between the channel and destination. The encoder applies some physical transformation to the transmitted message to make it suitable for transmission over the channel, encoding it as *transmitted signal*. When this resurfaces as *received signal*, the decoder performs the inverse transformation producing the received message, which finally arrives at destination (Shannon 1948). The ultimate goal of this encoding is to make reliable a noisy communication channel at a cost to be paid in terms of limitation to communication rates.

If we apply **SCM** to the original description of **AIP** given by Arrow, then it seems that the paradox somehow involves only these two possibilities:

(1) one transfers information through a single use of a noiseless channel (i.e., information flows at zero cost from source to destination);
(2) one keeps the information secret (i.e., the information transfer amounts to an infinite cost).

It can be shown quickly how to *extend* **AIP** through **SCM**, after renaming the components involved in the whole process. The process starts from an *idea*, that is *a payoff-relevant and privately observed piece of information*.[4] The *patent process* is the process transforming the idea in patent. The *patent use* is the process which extracts information from the patent; *intellectual property rights* (**IPRs**) play a role analogous to that of noise, Fig. 5.2. By **IPRs**, we mean the system which is in place in order to limit the use of an idea: from a communication point of view, **IPRs** act as an obstacle to information transfer.

On the other hand, information theory and cryptography may be considered "two sides of one tapestry" (Blahut et al. 1994). In 1949 – one year after his general model – Shannon shows how **SCM** may be displayed under a cryptographic model (Shannon 1949). The communication cost depends on the knowledge of a preliminary information, the *key*, which modifies the computational cost of decoding.

The key is used to modulate the level of noise in the channel. Inasmuch as the key protects information, it serves an analogous role as the legal infrastructure. So, it is hardly coincidental that **AIP** can be understood as exhibiting the same

[4] We adapt this definition from the one given for secret in Ganglmair and Tarantino (2014).

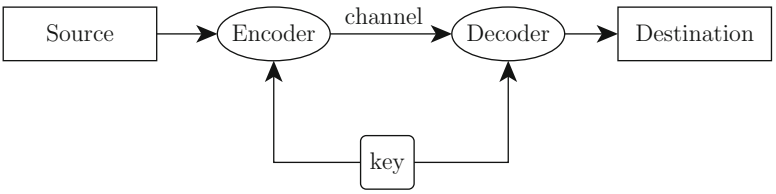

Fig. 5.3 **SCM** declined toward cryptography

cryptographic pattern: without the posses of the key, information transfer amounts to an infinite cost, with the key the information is available, see Fig. 5.3. Note well that *desynchronization* is at work here: the key is 'only' information needed to access to sensitive information. Buying the key is one thing and having the information is something else.

The provisional moral is that **IPRs** get around **H-AIP** by taking measure to render honest a potentially dishonest buyer. Yet, the buyer B buys the patent or the key *without knowing* the content of information. **E-AIP** is still looming.

5.6 AIP Through Zero-Knowledge Proofs

In 1985 László Babai introduced the notion of *interactive proof* based on the characterisation of the class NPTIME, i.e. the *non-deterministic polynomial time* complexity class (Babai 1985). This is the class of problems which can be decided in *exponential* time and whose solutions can be verified in *polynomial* time (Cook 1971). Significantly, the class NPTIME provides the archetypal way of generating *trapdoor one-way functions* for cryptography. One-way functions may be very informally presented as bijective functions such that their values are easy to compute, but whose inverse values are computationally intractable. Trapdoor one-way functions have the further restriction that they have an extra parameter, generally to be kept secret (the private key), which makes the inverse functions easy to compute. An example of a trapdoor one-way function is the product of two large primes, which is easy to compute but difficult to invert, up to knowing the trapdoor (one of the two factors).

Formalising a bit, the definition of interactive proof involves two parties (P and V) in the decision procedure of the language L:

$$\langle P(y), V(z)\rangle(x) = \begin{cases} 1 & \text{if } x \in L \\ 0 & \text{if } x \notin L \end{cases}$$

The pair of interacting algorithms P and V characterises the class of decision problems whenever the prover P requires exponential time resources and the verifier V polynomial time resources. Both receive the string x of L as input; P computes

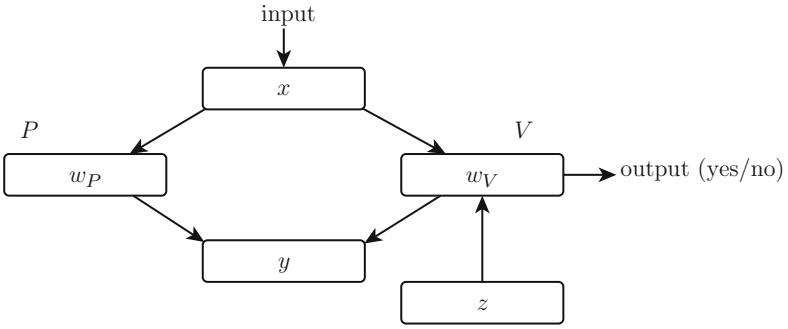

Fig. 5.4 Scheme of machines for interactive proving

a string y such that $|y| < p(x)$ for some polynomial p, and sends y to V. Then, V checks whether $y = f(x)$, where f is some computable function and z is used to ease the verification task (z is also called the "a priori knowledge" knowledge of the verifier) (Fig. 5.4).

In interactive proofs the two parties interact in order to solve a computational problem:

- P is called the *prover* since it is a computing entity (person, Turing Machine...) performing a 'computationally intensive' (difficult) task, which has to produce a summary of the information required to check the correctness of the result;
- V is called the *verifier*, which, thanks to information produced by P, can easily test its correctness.

It is worthwhile to stress that interactive proofs exist precisely in virtue of the asymmetric computational power between P and V: although V cannot access to the full work space of P, V is able to decide the problem (easily, i.e., in polynomial time) via the information provided by P. Note also that the output of the prover is required to be polynomial in size (even if the computation is not) in order to keep polynomial (in the size of x) the verification.

In 1986 Shafi Goldwasser, Silvio Micali and Charles Rackoff introduced the new concept of *zero-knowledge proof* (**ZKP**) by using that of interactive proof (Goldwasser et al. 1989). Their basic idea is that an interactive proof can be employed to show that one of the two parties is capable of performing a difficult task without giving any information on how the whole process can be performed.[5] The

[5]This typical example gives a good intuition to what zero-knowledge proofs look like in the real world. Alice and Bob are playing the game "where is Valdo": they have to find the real Valdo among a hundred of similar figures on the page of an illustration. How Alice can prove to Bob that she know where Valdo is without revealing his location? All Alice has to do is to take a large piece of cardboard (twice as large as the picture) with a small hole cut in the middle. She has to covers the picture with cardboard with the hole on the top of Valdo (while Bob is not looking), so that Valdo is lying behind the hole (Naor et al. 1999).

idea is that a **ZKP** is carried out by an interactive system (i.e., by using a *powerful* entity P) but after that V has decided $x \in L$ with the help of the prover P there exists a computable function S which can take the same decision without interacting with P:

$$\langle P(y), V(z)\rangle(x) = S(x, z).$$

Thus, for complexity theory an interactive proof is *zero-knowledge* if for any possible instance of the problem we have a polynomial time procedure to decide that instance, without having any information on the whole picture which remains a difficult task (requires the access to P).

The resolution of **AIP** we sketch here creates a bridge with **ZKP** by considering the *decision to buy* procedure. First of all, a formal setting must be introduced with respect to the very process of patent management. To this aim, we need to make available $\mathbb{L}$, a framework for language specification so conceived as to make possible to associate to any patent p a formal language $L_p \in \mathbb{L}$ of properties which the patent p verifies. At any moment, the buyer accesses the system in order to process requests of type $x \in L_p$; in fact, we may think of a buyer as an agent submitting a sequence of requests $x_1, x_2, \ldots, x_n$ and applying a deciding function $D(x_1, \ldots, x_n)$ which gives the determination to buy or not from the individual decisions (for the sake of simplicity, we can consider a global x which includes both the tests and the decision function).

Like in the **ZKP** computational setting, the prover P has enough resources (in terms of *computational power* or in terms of *knowledge* of the patent) to perform an interaction step: by accessing to the patent she can produce a trace of the property satisfied by the patent and pass this information to the verifier V which can then decide the input statement or continue with further interaction steps.

Here is a way to apply the zero-knowledge procedure to **IPRs**, so as to eventually circumvent **AIP**:

- the ***IPRs*** *owner* which wants to sell, is the prover P;
- the *buyer*, who has to decide to buy without direct access to the patent, is the verifier V;
- the *patent specification language* is a language L that the interactive proof can decide: it is given by the set of properties on which the two parties have an agreement and which empower the final decision of the buyer;
- P is capable of using the patent y to extract information under a challenge z from the verifier to assess some property x.

The interaction between V and P conveys *zero-knowledge* about the patent whenever there exists a decision procedure S such that for any property x and for any a priori information z used by the verifier to decide during the interactive protocol, we obtain the same outcome: this means that the very same reply is obtained as $S(x, z)$ without direct access to the patent.

Therefore, the important point to be made here is that there is no inspection of the information on the buyer's part *before* the purchase of it. By applying interactive

proofs to communication there will be benefits for both parties. On the one hand, the **IPRs** owner can reply to the requests of the buyer without giving direct access to the patent; on the other, the buyer can reach the final decision on the base of certified properties of the patent. There remains the task of specificating the *universal* formal language $\mathbb{L}$ through which *any* patent property can be specified. Of course, it is a formidable task (a formal language to give patent specifications as far as we know does not exist) but we have no strong reasons for thinking that it cannot be carried out. In real life the prover/verifier processes are not computational tasks (performed by Turing machines) but have to be managed in such a way that no information on the patent is carried through answers.

On the other side, other authors propose a **AIP** resolution based on step by step partial communication of the patent: the price determination (and therefore, at limit, the determination to buy) is obtained partially on the portion of patent disclosed (Horner and Skrzypacz 2016). In the zero- knowledge proof approach, the determination to buy can be obtained without having access to any single bit of information on the patent. Patent specification languages (yet) do not exist but especially in financial transaction, or in contract specification there are still several proposals (Jones et al. 2000).

5.7 Last Thoughts

In conclusion, **AIP** tells us the story of how information asymmetry becomes epistemically intractable when it concerns information itself (in all its forms). Patents and keys do not block **AIP** as a such, but **H-AIP**. As concerns **E-AIP**, we have argued that the notion of zero-knowledge proof may offer a *probabilistic* solution to **E-AIP**.

The intended moral for philosophers may be brought out by an analogy between **E-AIP** and Galileo's 'paradox' concerning the one-to-one mapping between natural numbers and their squares. Galileo's 'paradox gets internalized in Dedekind's *definition* of infinite set: a set is infinite exactly when it can be placed in one-to-one correspondence with a proper subset. The situation with respect to **E-AIP** seems similar: something is information exactly when one is willing to know it by performing either a blind or unnecessary action. **E-AIP** looms on honest people as a tax on their curiosity.

References

Akerloff, G. 1970. The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84(3):488–500.

Anton, J.J., and D.A. Yao. 2002. The sale of ideas: Strategic disclosure, property rights, and contracting. *The Review of Economic Studies* 69(3): 513–531.

Arrow, K. 1962. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, ed. J. Kenneth, 609–626. Princeton: Princeton University Press.

Babai, L. 1985. Trading group theory for randomness. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, STOC'85, 421–429. New York: ACM.

Blahut, R.E., D. Costello, U. Maurer, and T. Mittelholzer. 1994. *Communications and Cryptography: Two Sides of One Tapestry*. The Springer International Series in Engineering and Computer Science. Boston: Springer.

Cook, S.A. 1971. The Complexity of Theorem-Proving Procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, STOC'71, 151–158. New York: ACM.

Ganglmair, B., and E. Tarantino. 2014. Conversation with secrets. *The RAND Journal of Economics* 45(2): 273–302.

Goldwasser, S., S. Micali, and C. Rackoff. 1989. The knowledge complexity of interactive proof systems. *SIAM Journal on computing* 18(1): 186–208.

Horner, J., and A. Skrzypacz. 2016. Selling information. Journal of Political Economy 124(6): 1515–1562.

Jones, S.P., J.-M. Eber, and S. Julian. 2000. Composing contracts: An adventure in financial engineering (functional pearl). *ACM SIGPLAN Notices* 35(9): 280–292.

Merges, R.P. 1994. Of property rules, coase, and intellectual property. *Columbia Law Review* 94(8): 2655–2673.

Naor, M., Y. Naor, and O. Reingold. 1999. Applied kid cryptography or how to convince your children you are not cheating. In *Eurocrypt'94*: 1–12.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27: 379–423; 623–656.

Shannon, C.E. 1949. Communication theory of secrecy systems. *Bell system Technical Journal* 28(4): 656–715.

Stiglitz, J.E. 2002. Information and the change in the paradigm in economics. *The American Economic Review* 92(3): 460–501.

Thambisetty, S. 2007. Patents as credence goods. *Oxford Journal of Legal Studies* 27(4): 707–740.

Part III
Epistemology and Science

Chapter 6
Antimodularity: Pragmatic Consequences of Computational Complexity on Scientific Explanation

Luca Rivelli

Abstract This work is concerned with hierarchical modular descriptions, their algorithmic production, and their importance for certain types of scientific explanations of the structure and dynamical behavior of complex systems. Networks are taken into consideration as paradigmatic representations of complex systems. It turns out that algorithmic detection of hierarchical modularity in networks is a task plagued in certain cases by theoretical intractability (NP-hardness) and in most cases by the still high computational complexity of most approximated methods. A new notion, *antimodularity*, is then proposed, which consists in the impossibility to algorithmically obtain a modular description fitting the explanatory purposes of the observer for reasons tied to the computational cost of typical algorithmic methods of modularity detection, in relation to the excessive size of the system under assessment and to the required precision. It turns out that occurrence of antimodularity hinders both mechanistic and functional explanation, by damaging their intelligibility. Another newly proposed more general notion, *explanatory emergence*, subsumes antimodularity under any case in which a system resists intelligible explanations because of the excessive computational cost of algorithmic methods required to obtain the relevant explanatory descriptions from the raw data. The possible consequences, and the likelihood, of incurring in antimodularity or explanatory emergence in the actual scientific practice are finally assessed, concluding that this eventuality is possible, at least in disciplines which are based on the algorithmic analysis of big data. The present work aims to be an example of how certain notions of theoretical computer science can be fruitfully imported into philosophy of science.

Keywords Complex systems · Hierarchical modular descriptions · Modularity · Antimodularity · Explanatory emergence · Scientific explanation · Computational complexity

L. Rivelli (✉)
Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA), University of Padova, Padua, Italy

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_6

6.1 Introduction: Modularity, Explanation, Philosophy of Science and Computer Science

My aim in this work is to show how philosophy of science, and specifically the philosophical reflection on the kinds of explanation typically employed in certain special sciences, should be concerned with the computational complexity of certain algorithmic tasks regarding the detection of a modular hierarchical structure in the observed systems.

First, I will focus on *hierarchical modular descriptions* of complex systems and try to show that they are at the core of mechanistic and functional explanations, two kinds of explanation widely employed in science. I will also try to show how certain computational limitations and constraints affect the production of hierarchical descriptions of complex systems, and how this can hinder scientific explanations which are based on descriptions of such a kind.

Starting from seminal works by Herbert Simon, I will try to sketch a philosophical framework for the definition of the notion of *modularity* in general, showing how this concept is applicable to a wide class of descriptions of phenomena, and how modular descriptions constitute a necessary ingredient of both mechanistic and purely functional explanations, as analyzed by current philosophy of science. I will try to elucidate the main kinds of modularity, the typical properties of modular systems and the relationship between structural, dynamical and functional modularity. I will then show how mechanistic and functional explanations are based on modular descriptions, and proceed highlighting that the hierarchical nature of *multi-level* modular descriptions bears on the *intelligibility* of the associated explanations: availability of the sole *low-level* description of the elementary parts of a complex system and their relations, with the lack of a corresponding multi-level hierarchical description comprising also structural and functional high-level macro-modules, can severely limit our understanding of the functioning of large enough complex systems, due to the huge amount of detailed information provided by the low-level description, an amount of information which could possibly result overwhelming with respect to human cognitive capacities. Availability of higher-level descriptions can instead allow us to render more intelligible, by a simplifying coarse-graining, the structural and functional organization of the observed system: a multi-level description renders the system intelligible by allowing a fine-tuning of the amount of information conveyed by the description, according to the observer's needs and capacities.

But, how to obtain a hierarchical modular description of a complex system? In recent years, in sciences such as systems biology, attempts at bottom-up reconstructions of the hierarchical structure of mechanisms starting from data already available by means of high-throughput methods of low-level analysis of the system have become popular. In these cases, due to the enormous size of the original datasets, constituted by a myriad of interactions between the basic elements of the system (for example protein-protein interactions), it is typical to recur to automated algorithmic methods, for the production of hierarchical modular descriptions. I will focus on

methods for the detection of modularity in *networks*, because a network (or graph) is the typical kind of structure employed to represent the original dataset of low-level interactions.

Now, it turns out that algorithmic detection of hierarchical modularity in networks is a task plagued in certain cases by theoretical intractability (NP-hardness), and in most cases affected by the still high computational complexity of most polynomial-time approximated methods. This renders the reconstruction of the hierarchical modular structure of these systems highly problematic when the system reaches a certain scale, leaving us with the low-level description of the system as the only available description, a description which, as highlighted above, could easily turn out to be scarcely intelligible, and thus not very useful from an explanatory standpoint. This circumstance has prompted me to put forth a new notion, *antimodularity*, which consists in the impossibility to algorithmically obtain a hierarchical, multi-level, modular description fitting the explanatory purposes of the observer, for reasons tied to the computational cost of typical algorithmic methods of modularity detection, in relation to the excessive size of the system under assessment and to the required precision. As expected, the occurrence of antimodularity, by preventing the production of intelligible and valid hierarchical descriptions of the systems, hinders both mechanistic and functional explanations by damaging their intelligibility. I then propose another more general notion, *explanatory emergence*, which subsumes antimodularity under any case in which a system resists intelligible explanations because of the excessive computational cost of algorithmic methods required to obtain the relevant explanatory descriptions from the raw data.

Finally, the possible consequences, and the likelihood, of incurring in anti-modularity or explanatory emergence in the actual scientific practice are assessed, concluding that this eventuality is possible, now or, possibly, in the future, at least in disciplines which are based on the algorithmic analysis of big data, even if some factors could render occurrence of antimodularity not very evident in the scientific literature.

In light of the above anticipation of what will follow, it is clear that, in addition to its central specific theme, which is about hierarchical modularity, algorithmic methods for its detection and their limits, and the importance of the obtained hierarchical modular descriptions for the scientific explanation of complex systems, a non secondary general aim of this work is to show how some notions of theoretical computer science can be imported and fruitfully employed in philosophy of science. This, it seems, is not a widespread practice: with the exception of the basic notions of computability (the Turing machine model, the halting problem), not many of the main results of theoretical computer science are often taken into consideration by philosophers of science. In particular, the notion of computational complexity has more or less been neglected in the philosophy of science literature: usually, the philosopher is content with knowing or claiming that something is computable, often ignoring the possible costs, in terms of time or memory space, of the required computational task. In actuality (unless P turns out to be equal to NP, an eventuality which seems unlikely), the costs of certain computational tasks have

been mathematically proved to be unavoidably too high for them to be carried out, at least for certain inputs. This is the case of any algorithmic method whose purpose is to solve exactly the problems which have been proved to belong to the *NP-hard* class, the so-called *intractable* problems. Such problems, to be solved exactly by an algorithm, require a time which is an exponential (or *more* than exponential) function of the input size, and as a consequence, their algorithmic solution could, in the worst cases and for sufficiently large inputs, take times longer than the actual age of the universe to come to completion: even if those problems are *in principle* computable, they are not feasibly computable *in practice*, and we must resort to approximate methods. Usually, these approximate methods turn out requiring a run time which is a *polynomial* function of the input size, and as such are considered tractable. But, as is well known in computer science, not every intractable problem is approximable in polynomial time with sufficient precision. And, when complete approximability is precluded, even the approximate, polynomial-time algorithms that manage to yield sufficiently precise results could turn out to be too computationally heavy: they are executable in polynomial-time, but with a polynomial of a quite high degree. For algorithms which solve approximately in polynomial time an intractable problem, there is often a trade-off between the precision of the algorithm and its running time: the higher the required precision, the slower the execution time. As we will see, this turns out to be the case for most algorithms for modularity detection.

A second aim of this work is then to propose the difficulty of algorithmic detection of modularity as an example of why and in what circumstances philosophy of science should be concerned with certain more advanced themes of computer science such as computational complexity: in other words, an example of how computer science can fruitfully inform philosophy of science.

6.2 Modularity

Although modularity appears to be a basic and ancient notion, modern philosophical reflection on modularity has only begun in the second half of the twentieth century, with the especially relevant contribution of Herbert Simon under the form of his notion of *hierarchical nearly-decomposable* systems,[1] that is, systems that can be seen, at least as a first approximation, as recursively, hierarchically decomposable into sets of cohesive, partially independent subsystems. This is the basic idea which inspires my proposal on modularity. In this work, I will examine a possible conception of modularity in complex systems, and explore the consequences of its presence or of its absence (a circumstance which I call *antimodularity*) on the explanation of the structure and behavior of such systems.

[1] See the seminal (Simon 1962).

Embracing a widely epistemic stance, along the lines of Cory Wright and William Bechtel's *epistemic* position on mechanistic explanations,[2] I consider scientific explanations as *epistemic devices*, based on *descriptions* of phenomena, related to human *communication*, and requiring at least a minimum degree of cognitive *intelligibility*. Accordingly, I am interested in defining modularity as a feature not of actual, real systems, but *of their descriptions*, a feature which, if present, allows for certain comprehensive types of explanation.

6.2.1 Modularity in Complex Systems

Proceeding along the lines expressed above, I will consider modularity in complex systems. A complex system is to be intended here simply as *a set of interrelated parts*. I informally define the property of *modularity* in complex systems as the possibility for a system of this kind to be described as a set of loosely related *modules*, that is, a set of well-defined, cohesive subsystems, with internal parts highly interconnected, each subsystem partially independent from the external context, being it only weakly connected to other subsystems. In other words, modularity of a complex system basically manifests itself as the possibility of *decomposing* the system into recognizable, sufficiently defined and persistent subsystems, that is the modules, *each module composed of parts which are more strongly related to each other than to parts belonging to other modules*.

I extend this view of modularity to that of the full *hierarchical modular description* of a system in terms of "higher" and "lower" levels of description, each of which is constituted by modules, and where, except for the lowest level, each module at one level is a *macromodule* that can in turn be seen as internally characterized by a modular organization of *micromodules*, and so on recursively. In line with the essence of an epistemic view, all of this concerns descriptions, not sets of real-world objects.

The point to highlight here is that the whole hierarchical modular description turns out to depend, due to the definition itself of modularity, on the observer's *choice* of a specific significant *relation* between the elementary parts of the system, and this precisely because of the way the concept of *module* is defined: a module is a subset of the parts of a whole that are *related* to each other in a stronger way than how they are related to parts external to the module they are in. Recognition of a subset as a module requires thus that a *relation* between parts is taken into consideration first, and, depending on which specific relation is considered, the identifiable modular structure can change.

[2]This *epistemic* position is usually opposed to an *ontic* conception of causal explanations, which considers the actual mechanism *themselves* as their *explanations*. See Bechtel and Abrahamsen (2005) and Wright (2012).

This definition of hierarchical modularity presupposes that a complex system is composed of distinguishable, related, elementary parts, and this in turn is due to the choice of an atomic, elementary description of the system: the choice of the set of *parts* and that of the *relation* holding between them amounts to the pragmatic *choice*, on the part of the observer, according to her interests, of what could be called the *basic description* of the system. This highlights a *pragmatic* component which, it seems to me, is constantly present in explanation. In actual science, a "natural" lowest-level description is often suggested by the physical properties of the system combined with the researcher's interests: for example, in biology a tissue is naturally described as composed of cells, or a cell is naturally described as composed of interacting macromolecules. The point to stress here is that hierarchical modularity is *relative* to such a choice, depending especially on the choice of the *relation* holding between the system's elementary parts (which usually is a less constrained choice than that of the parts themselves): for example, the same set of individuals (a population) could be seen as interrelated by a relation of parenthood or by a relation of friendship, and the modular descriptions would vary accordingly.

6.2.2 Modularity, Decomposability, and Economy of Description

As said, modularity manifests as the possibility of *decomposing* a system[3] into recognizable, sufficiently well-defined subsystems, each one composed of parts which are more strongly related to each other than to parts belonging to other modules or, in general, to the extra-module environment. It is the presence of these variations in strength of the relation holding between couples of parts of the system, which allows for the *recognition* of modularity: if all parts of the system were fully connected to each other with the same intensity, modules would not appear, because a module is (informally) defined as a subsystem whose strength of connection with the rest of the system[4] is *lower* than that of the connection between the module's internal parts. This conception is quite similar to the notion of *near-decomposability*, originally conceived by Herbert Simon[5]: near-decomposability allows the original system to be represented as a set of loosely connected subsystems (the modules) with a higher internal cohesion, and this decomposition can be reiterated until a full hierarchical description is obtained. The crucial point is that the original system, composed of its elementary parts, is thus describable in a *high-*

[3] With "system" here I mean a *description* of a system. In what follows, I will often use the term "system" simpliciter to mean its standard description, usually the *basic description* (see Sect. 6.2.1).

[4] A connection with the rest of the system effected obviously by means of individual links going from nodes internal to the module toward *nodes* belonging to other modules.

[5] See Simon (1962).

level manner, under the form of *another* system whose *single parts* correspond, each one, to one *module* of the original system: this is a form of *aggregation*, in which a single part of the high-level description comes to represent the *aggregated value* of the dynamics of a whole module. In general, due to aggregation, the high-level description turns out to be coarser-grained and thus *simpler* than the low-level one, because, in the former, entire *groups* (the modules) of low-level parts are represented as *single* high-level parts, and so the parts of the higher level description are fewer in number than the low-level ones. This way, the high-level description appears usually more economical and perspicuous than the original one: the smaller number of interrelated parts in the high-level description renders it more graspable by the observer from a cognitive standpoint.

6.2.3 Structural and Dynamical Modularity

It is easily conceivable that modularity can not only concern the *structure* of a system, but also its dynamical functioning: dynamical modularity can be conceived as modularity of the *process* occurring in a system, and this can be seen as the fact that some changes of state in some of the parts of the system are *temporally* related one another *more closely* than they are related to other changes of state occurring in different parts of the system at different times. The relation between structural and dynamical modularity turns out to be not always a simple relation. Structural and dynamical aspects can be associated but also decoupled, albeit in most cases of dynamical systems their modular physical structure *induces* a form of modular dynamical functioning, given that in dynamical systems the dynamics is conducted *on* the system's predefined structure and it is thus constrained by it: just think of an electric circuit, where the structure of the connections between components determines the dynamical flow of the electric currents. In general, the more structurally related two elements are, the more easy and/or frequent is the possibility of them interacting by exerting some kind of influence which determines the change of state of the other, and thus the more *dynamically* related they are.[6] This is reflected in the presence of a *temporal decoupling* between the dynamics occurring *inside* modules and those occurring *between* modules: due to the stronger relations between elements belonging to the same module, intra-module interactions tend to occur more frequently than inter-module ones. It is this time-scale decoupling that allows the observer to (approximately) *isolate* each module's dynamics from the global dynamics of the system, and to study each module separately, considering the system as approximately decomposable (that is,

[6]Exceptions to the common association between structural modularity and dynamical modularity can occur when the system's parts have highly non-linear responses to inputs: in that case, even an interaction along a structurally weak connection between two parts can induce a disproportionately strong response on the receiving part, due to the non-linearity of its input-output function.

nearly-decomposable): at the time scale of the internal processes of a module, the external environment, which is orders of magnitude slower in its dynamics, appears approximately *still*, and can be considered as constituting a static external condition, allowing the study of the internal module dynamics separately from the external influences.

6.2.4 Aggregation in Dynamical Systems

In certain cases, when describing the whole system as constituted of interrelated modules, we can coarse-grain it by re-describing it in an aggregated manner, in which to each module is substituted some variable which in a way or another represents the whole module's dynamics. For example, in thermodynamical systems, we can substitute the mean temperature of the module to the actual internal temperature distribution of the module itself. This is a form of functional modularity deriving from Herbert Simon's notion of near-decomposability, which consists in the *aggregation of variables* of the dynamical model of a system[7]: it is a form of near-decomposability *of the mathematical model* describing the system's dynamics, which allows the production of a simplified (i.e. with less variables) dynamical model which constitutes a simpler redescription, modulo a certain accepted amount of approximation, of the system's dynamics. A more detailed exposition of this kind of modularity is outside the scope of this paper, but it must be highlighted here that the search for a suitable, even approximate way to aggregate the variables, even for completely linear dynamical models, has turned out to be plagued by computational intractability (NP-completeness and NP-hardness), as showed in Winker (1992) and Kreinovich and Shpak (2006).[8]

6.2.5 Modularity and Explanation

It appears that modularity is linked with scientific explanation in various and fundamental ways.

The mathematical formula employed by the dynamical model representing the aggregate dynamics of a nearly-decomposable system is *simpler* than the formula of its original dynamics, and this means that aggregability produces an economy of description. Since a scientific explanation of the system's dynamics (at least a deductive-nomological type of explanation[9]) would surely employ this formula, modularity (in the form of aggregability) produces an economy of explanation.

[7] See Simon and Ando (1961).

[8] See also Kreinovich and Shpak (2008).

[9] See Hempel and Oppenheim (1948).

I argue also that modularity of a system's description, as expounded above, is necessary for *mechanistic explanation*, a model of scientific explanation which has been since the 1990s the object of two main lines of philosophical inquiry: one line by William Bechtel and his collaborators, and another by Carl Craver and his colleagues.[10] In the account of William Bechtel and Adele Abrahamsen, basically a mechanism is seen as "a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena."[11] The point to stress here is that there is a functional view involved: the global function, that is the phenomenon produced by the mechanism, which represents the explanandum, is explained by describing the organization and interactions of the parts which, by means of their dynamical "orchestrated" functioning, produce the phenomenon. What is needed is then to first identify the parts and operations involved in its production. To this aim, the system as a whole must be subject to two operations: *structural decomposition*, which yields the set of elementary parts, and *functional decomposition*, which identifies component operations. A third, desirable phase is *localization*, consisting in linking the elementary parts with the operations they perform. It seems to me this whole conception of mechanisms could be easily rephrased in terms of *modularity*, along the lines of the view which I have sketched above. The result of functional, structural decomposition and localization is what I have called the *basic description* of the system: the identification of the basic, lowest level parts which the observer has chosen to identify and of their interactions.

The resulting low-level kind of explanation is not always the most desirable, and, as Bechtel and Abrahamsen highlight, it is important that a whole *hierarchy* of mechanisms be considered, and that explanation be *multilevel*: circumstances external to a given mechanism can be seen as larger overarching mechanisms, while components of a mechanism can be seen as mechanisms themselves, recursively composed of subparts.

A mechanistic explanation tries to answer questions about "how" a phenomenon is brought about by showing the way the complex dynamical functioning of a set of interacting parts produces the phenomenon. The same questions can be answered to, also just from the *functional* point of view: this conception, *functional analysis*, has been notoriously advanced by Robert Cummins, starting from his seminal Cummins (1975). Similarly to mechanistic decomposition, functional analysis consists in the recursive decomposition of the global phenomenon, taken as the overall function to be explained, into its component subfunctions. Seen from an explanatory point of view, the overall function of a system is explained by means of the organized contributions of its subfunctions, which are executed in a programmed activity. This

[10] The two corresponding seminal works are Bechtel and Richardson (1993) and Machamer et al. (2000). The line led by William Bechtel proposes the so-called *epistemic view* of mechanisms, which I also endorse (see Sect. 6.2). This is opposed to the *ontic* view of mechanisms, mainly supported by Carl Craver. See Wright (2012).

[11] Bechtel and Abrahamsen (2005, p. 423).

position is quite close to a computational view, and it is completely compatible with it. Actually, Cummins' proposal is the prototype of the typical explanation of cognitive psychology, which mostly consists of functional explanations, often in the form of *computational explanation*, that is, the exhibition of a computer program able to produce the cognitive phenomenon to be explained. This is a typical form of so-called *role functionalism*, in that here the concept of function is the concept of a partial *role* fulfilled by a subsystem in order to bring about, in interaction with all the other functions fulfilling each one its role, the whole functioning of the overarching system.

A point to highlight here is that the recursive functional decomposition until a full hierarchy is obtained is the strategy to be sought for when pursuing functional explanations of complex systems. This hierarchical functional decomposition is required for the production and comprehension of a functional explanation: the reason is that the *role* that can be attributed to any subfunction is defined in relation to the higher-level capacity (the explanandum) that the subfunction contributes to bring about together with all the other subfunctions at the same description level: a function is thus recognized as such in relation to all the other functions at its level and in relation to the higher-level overall function. In other words, the role of a function must be understood in relation to all the functions of the system (or of the subsystem) which is the object of consideration. But the mind of the researcher must be *capable* of obtaining this "holistic" understanding in order to produce a functional explanation: otherwise there would be no *functional* explanation at all. This task is greatly eased by the possibility of representing the system as a hierarchy of subsystems, and thus as a hierarchy of functions: in this case a function at any level has to be put, to be understood as a function, only in relation to the overall function of the subsystem and to the other sibling functions composing the subsystem. Given that each subsystem is smaller than the whole system, this tends to greatly simplify the task of functional analysis.

In general, high-level modularity should allow for a form of *coarse-graining*, understood as the operation of taking a complex system represented as a set of many parts, partitioning this set into disjoint subsets, and considering, in place of the original system, *another* set in which each *part* corresponds to one of the disjoint *subsets*. This is basically the same operation, whether effected on sets of variables of an equation, as in aggregation, or on a network,[12] where the original representation can be substituted with a network with fewer nodes, or in the case of functional and mechanistic explanations, where a group of interacting parts or actions can be seen as a whole function, or mechanism, and a group of mechanisms can be seen as a single super-mechanism, whose parts are the simpler single mechanisms. In each of these cases economy of description is achieved by the coarse-graining allowed by high-level modularity, and arguably, understandability of the explanation is greatly eased, because the high-level, coarse grained description is constituted of fewer parts, and thus is *simpler* than the basic description, which is constituted instead of all the numerous elementary parts and their interactions.

[12] See Sect. 6.2.6.

Considerations of economy or intelligibility aside, modularity is *necessary* to produce certain types of explanation. As we have seen, Robert Cummins' analytical functional explanatory strategy explicitly pleads for a hierarchical decomposition of the system's functioning, in order to explain it. Of course, this decomposition is possible just in the case some form of functional high-level *modularity* is present in the system, that is, when the high-level modules to be sought for can legitimately be considered functional modules.

Similarly, for a mechanistic explanation, at least in the conception put forth by William Bechtel and his group, a conception which does not consider mechanistic explanation as merely reductionistic, it is desirable that the explanation be *multilevel*, and this corresponds to a hierarchical functional-mechanistic description of the system. Embracing an *epistemic* view of explanations, these authors quite naturally highlight also the importance of the cognitive *intelligibility* of explanations, and this can be achieved by the hierarchical modularity of the descriptions employed in explanations: *hierarchical modularity allows for multilevel explanations*, which certainly enhance comprehension.

In general, given an appropriate hierarchical modular decomposition, a system can be described at any desired level of description, with different results on the intelligibility of the explanation: the more abstract, coarse-grained levels allow for a very simplified explanation, which usually induces better understanding, while the choice of proceeding down to more detailed lower levels enhances the information on the system conveyed by the explanation, possibly at the cost of understanding, for too much detailed information could overload the observer's cognitive system. The most detailed possible explanation is the one which describes the system in terms of its elementary parts, but, in many cases, the sheer amount of information contained in such a description can hinder its intelligibility. The best explanation is then the one which minimizes the trade-off between understandability and detail provided, but this fine tuning of the explanation *requires* the possibility of describing the system at several different hierarchical modular levels. Thus, *absence of hierarchical modularity hinders mechanistic and functional explanation*.

6.2.6 Detection of Modularity in Networks

Algorithms for *modularity detection* are procedures which, given a basic description of a complex system in terms of elementary interrelated parts, try to produce a *hierarchical modular description* of the system. I consider here, specifically, algorithms for modularity detection in *networks*, because network models constitute a typical way of representing complex systems, especially biological systems, in recent research.

Fig. 6.1 A network

A *network* can in general be seen as a set of parts, its *nodes*, connected to each other in various ways through *links* or *edges*[13] (as in Fig. 6.1).

There are several, not incompatible, possible forms of modularity in networks. Here, I take into consideration modularity understood as the presence of *community structure*, based on the conception of modules as cohesive subsystems weakly connected one another, and modularity understood as the recurrence of *network motifs*, coinciding with the idea of modules as repeatable standard functional parts.

6.2.6.1 Community Structure

Community structure, as exemplified in Fig. 6.2, is the property of a network to be composed of *communities*, that is, roughly stated, subsets of nodes whose internal nodes are more densely (or intensely) connected one another than how densely (or intensely) they are connected to nodes belonging to other subsets. This property, initially proposed by the classic works of Mark Newman and Michelle Girvan,[14] reflects quite directly the definition of modularity I proposed above and Simon's original definition of near-decomposability, as applied to networks.

In Newman and Girvan (2004) a measure of the quality of a modular description, called "modularity", or "Q", has been proposed, and subsequently this measure has spread in the literature as a paradigmatical reference for the task of community

[13] See Newman (2003).

[14] Starting from Girvan and Newman (2002).

Fig. 6.2 A network with community structure. In this picture, discs surround the communities, which show high density of intra-module links, while external, inter-module links, are more sparse

detection. In general, community detection can be seen as the task of maximizing the value of Q, that is, the task of finding, among the exponential number of possible partitions of a network, the one which has the highest value of Q: this partition is the one which best describes the *actual* (if any) community structure present in the network under observation. This stems from the fact that Q is defined in a way that reflects the very notion of modularity in general, on which the notion of community is based: that is, the idea that modules are cohesive subsets, highly connected inside, and sparsely connected one to the other. Accordingly, the basic intuition behind measure Q is that the density of interconnections between the nodes inside a community (inside a module) must be significantly higher than the density of interconnections between the corresponding nodes in a randomized version of the network under observation in which nodes are connected at random, but which respects the degree distribution of the original network.[15] So, a high value of Q for a modular description means that the modular description detects communities which are much more densely connected internally than they would be if the nodes in the network were linked at random, and thus that the description really catches a significant modularity which highly differs from the "null" model of the network (its randomized version, taken as a benchmark), where distribution of links on the network is expected not to show any significant modular structure. In other words, a high value of Q means that a modular description really catches a modular

[15]The *degree* of a node is the number of links to which it is connected. In the randomized version, each node, even if possibly connected to different nodes than in the original network, has the same degree of the corresponding node in the original network.

structure which is *actually* present in the network, and that cannot be due to random fluctuations in the density of links. It is clear then that the goal of a community detection algorithm is that of *maximizing* Q.

6.2.6.2 Network Motifs

Another form of modularity in networks, originally proposed by Uri Alon, Ron Milo and others,[16] consists in the presence in a network of recurrent patterns of connected nodes, the so-called *network motifs*. This is a form of modularity which exemplifies the idea that a module can be seen as a subsystem recurring in different copies in different parts of the system, and that each type of recurring subsystem performs a basic standard function. If we understand the network as the fixed structure *on which* a dynamical process can occur (as for example in the case of electronic, or boolean circuits), then a network motif can be seen also as a functional module, especially in directed networks, where each node can be seen as receiving inputs from links pointing to it, and producing outputs towards links stemming from it. The most common network motifs are simple patterns of interconnection which can implement, in directed networks, simple functions which are typically studied in control systems theory, such as feedback and feed-forward loops (see Fig. 6.3).

In general, given that functional and structural modularity, even if conceptually distinct, are often related, methods for automated detection of modularity in networks, which apply to the network *structure*, could, if applied to a network representation of a dynamical system, yield an immediately functional modular description. And motif modularity can be easily combined with community modularity, in that motifs can be often seen as basic building blocks making up communities, which in turn represent higher-level functions: it is conceivable (see Fig. 6.4) that modularity detection can proceed by first locating recurring motifs, a phase after which a motif can be considered a standard module performing a certain function recurring in different parts of the network (the same way a type of standard electronic component, for example a transistor, occurs and performs a standard function in several parts of a circuit). We can subsequently recognize *communities* composed *of* motifs, which could represent higher-level functional modules, and so on, until a full hierarchical modular representation of the network is obtained. This would be a structural and functional decomposition of the system: under the functional aspect, community structure reveals the high-level functional patterns of interconnection *between the modules*, while motif modularity reveals the internal subfunctions which make up each module.

[16] As e.g. in Shen-Orr et al. (2002).

Fig. 6.3 Two typical network motifs. (**a**) Feedback loop; (**b**) Feed-forward loop

Fig. 6.4 An example of possible structural/functional hierarchical modular levels of description in modular networks. (**a**) nodes; this is the *basic level*. (**b**) motifs. (**c**) communities, composed of motifs. (**d**) the whole network. The vertical arrow points from lower to higher levels

6.2.6.3 Computational Complexity of Modularity Detection in Networks

Detection of modules in networks is effected through algorithmic methods, which of course are best performed by computational devices.

As already highlighted in Sect. 6.2.5, hierarchical modularity is especially important for the intelligibility of functional and mechanistic explanations of the system. It turns out that most of the best algorithms for the detection of the hierarchical modular structure of networks are computationally highly demanding, and there are also, in certain cases, theoretically established limits on the feasibility of the detection of specific kinds of modularity: many of these tasks, in their exact form, are computationally *intractable*.

For what concerns community structure, it has been proved in Brandes et al. (2008) that the algorithmic search for the most accurate modular description of a network (the so-called Q *optimization* task[17]), which is the paradigmatic method for community detection, is hindered by an insurmountable computational time complexity: in its decision variant, the task is *NP-complete*. As a consequence, most algorithms for detection of community structure implement approximations of this optimization task. But it turns out that most of these algorithms for simply *approximating* the optimal detection of community modularity in networks are themselves quite computationally intensive, even if they run in polynomial time[18]: some proposed algorithms for community detection can have complexity $O(n^4)$ or even $O(n^5)$, $O(n^6)$ or $O(n^7)$.[19] Most or the more widespread algorithms in use are $O(n^2)$ or $O(n^3)$, although some recently proposed approximate solutions running only in linear time have turned out to be quite precise.[20] However, even many of these faster methods, which manage to run in linear time on sparse networks, become costlier, running at least in $O(n^2)$ in case the input network is denser, as showed in Papadopoulos et al. (2011).[21] In general (with some exception), due to their approximate nature, often making use of stochastic sampling methods, such algorithms suffer a trade-off between precision and speed. In certain cases, the algorithm's precision is theoretically limited, as for the algorithm proposed in Brandes et al. (2008), whose precision with respect to the optimal detection of community structure is at best a factor of 2. Good et al. (2010) shows that the dominant measure Q of quality of a modular description is a "degenerate" function, in that it presents many local maxima by which approximate algorithms, which should seek for the *global* maximum, could be induced in error. This would explain the apparent ease with which the NP-hard task of maximizing Q can be

[17] See Sect. 6.2.6.1.

[18] As surveyed in several articles, e.g. Danon et al. (2005), Orman et al. (2009), Yang et al. (2010), Papadopoulos et al. (2011), Orman et al. (2011), Plantié and Crampes (2013) and Chakraborty et al. (2016).

[19] See for example Papadopoulos et al. (2011).

[20] For example Blondel et al. (2008).

[21] Table 1, p.529.

approximated in polynomial time. The article concludes: "[...] a cautious stance is typically appropriate when applying modularity maximization to empirical data. Unless a particular optimization or sampling approach can be shown to reliably find representative high-modularity partitions, the precise structure of any high-modularity partition or statistical measures of its structure should not be completely trusted".[22] A variant of Q, a quality measure for community detection in *bipartite* networks, introduced in Barber (2007), has been proved in Miyauchi and Sukegawa (2015) to be NP-hard as well.

Algorithms for detection of network motifs are also plagued by a high computational cost. The first theoretical reason is that an essential step in the task of exact detection of motifs consists in solving the problem of *subgraph isomorphism*, which is notoriously *NP-complete*.[23] The second is that the number of possible motifs is exponential in the number of nodes that compose a motif, and so the search for motifs larger than 5 nodes is very difficult. Approximated algorithms, as expected, are still quite computationally heavy, and they suffer from a precision/speed trade-off. According to Wong et al. (2012), which is a survey on widely used motif detection algorithms, in real-world tests on well-known biological networks many algorithms cannot cope with large networks: some of the best algorithms employ, just to search for 8-nodes motifs, times of the order of many months, while they result prohibitively heavy for larger motifs.[24]

In general, it appears that the algorithmic detection of network modularity is affected by a trade-off between complexity of the task and dependability of the modular description produced, and for this reason the identification of approximate but still acceptable hierarchical descriptions could result unfeasible, depending on the observer's needs, for systems of sufficiently large size: the time complexity of the task combined with the system's size could render the detection of hierarchical modularity a process too slow to be of practical use.

6.3 Antimodularity

Given the above observations about the high computational cost of the algorithmic detection of modularity, I propose to define the property of *antimodularity* in general, as *the impossibility of obtaining, by means of algorithmic modularity detection, an explanatorily useful and valid hierarchical modular description of a system*. More precisely, a system shows antimodularity when its most feasible and faithful hierarchical description, yielded by algorithmic means, is too approximate to be a useful high-level explanation of the system anyhow, or it is even

[22] Good et al. (2010), p. 10.

[23] See Garey and Johnson (1979).

[24] See Wong et al. (2012), p. 9, Table 4, and Sect. 6.5.

completely invalid, in the sense that the (presumed[25]) dynamical behavior of the obtained hierarchical description drastically diverges from the dynamics of the original description, rendering the hierarchical description so obtained useless for explanatory purposes. In these cases, the only possible *valid* hierarchical description is the *basic description*: in other words, *antimodular* systems are systems which, intuitively, can be explained by modular descriptions at *one* level only, the level of their elementary, finer parts, which I called the *basic description*.

Antimodularity is due to failure of the application of algorithmic methods for modularity detection, and this in turn can be possibly blamed on two conditions:

1. No intermediate-level modularity can be reasonably supposed in the system, given its basic description. That is, roughly stated, the system so described is actually *not* modular at any level higher than its chosen basic description. In this case antimodularity is intrinsic to the given basic description, no matter how accurate the algorithm for its detection is. I call this condition *intrinsic antimodularity*. This situation can occur when there is a more or less uniform distribution of the strength of the relationship between parts across the basic description of the system, and so the criterion for distinction of modules cannot be applied even in principle: modularity, relative to that chosen relationship between parts, is actually, objectively, absent.[26]
2. *Regardless of the fact* that an actual modular structure is present in the system's basic description or not, antimodularity arises because, given the *high number of parts* composing the system's basic description, *and the precision required by the observer*, the modularity-detection algorithm ends up being too *computationally expensive* to be brought to completion, either because it is computationally hard, or, although formally not intractable, because it is too computationally costly to be brought to an end anyway, in that it is polynomial time, but with a too high degree. This circumstance leaves the observer with the sole possibility of resorting to a description, obtained with a more feasible algorithm, which is however too approximate for her explanatory purposes. I call the occurrence of such a circumstance simply *antimodularity* (of course, intrinsic antimodularity is a case of antimodularity).

[25] A check of the validity of the modular model would involve its use as a model for a computer simulation, in order to compare its behavior with that of the actual empirical system. But not every explanatorily useful description can be immediately used as a dynamical model for a simulation: in certain cases, a high-level description is able to elicit comprehension of a system's functioning without providing the necessary details for its implementation as a dynamical model that can be directly put to test in a simulation run. A typical example would be one of the high-level functional block diagrams typically used in cognitive psychology to describe general mental functions: their actual implementation would constitute a computational explanation able to be used as a computer simulation of one or more of the main human mental functions: we obviously probably still lack this explicit implementation of an intelligent system. But we could in certain cases still be able to discern completely invalid functional diagrams from more plausible ones.

[26] Intrinsic *functional* antimodularity can occur even if there is an apparent *structural* modularity, because in certain cases structural and functional modularity do not coincide. This can happen when the system is highly non-linear: the non-linearity of the input-output functions of the nodes can make even weak connections between them trigger intense responses, preventing this way the temporal decoupling between what appear as structural modules.

The reason behind this antimodularity/intrinsic antimodularity distinction is that, while antimodularity could in some case be eliminated by improving the modularity-detection algorithm, intrinsic antimodularity would still hold in any case, being not due to the inaccuracy or to the computational cost of the algorithm employed, but to an objective feature of the chosen basic description of the system.

It appears then that modularity detection could, in sufficiently large systems, be actually *prevented* by problems of *computational cost*, or even computational hardness, so a system can be *pragmatically* considered antimodular, even if in principle it possesses some modularity, which, however, we are practically unable to detect. An *antimodular* basic description of a system does not possess, *at least as far as we can know*, any valid high-level modular redescription which we deem sufficiently useful in order to explain the system.

The *pragmatic* aspect of antimodularity, anyway, should not be downplayed as *merely* pragmatic: it is a pragmatic impossibility to bring to completion in a feasible time a computer program, but, especially when the computational hardness of an algorithm has been mathematically proved, this pragmatic hindrance becomes something more compelling, assuming the cogency of a logical law: unless P turns out to be equal to NP (and this seems unlikely), there cannot be any hope of rendering the *exact* task, which has been proved to be computationally hard, more computationally feasible. Unless $P = NP$, it does not matter how we try to improve the exact algorithm, or improve the power of the system on which it runs: its execution time will, at least in certain cases, always overcome any possible improvement in speed. It must be stressed that we are talking here of the *exact* task, which is provably NP-hard. But such a task can be probably *approximated* in a more reasonable time. Even in this case, however, the trade-off between speed and accuracy, which is typical of approximated algorithms for modularity detection, associated with the high number of parts of some complex systems, and the precision required by certain types of explanation, (as a typical case, mechanistic explanation), could make the produced modular description excessively approximate for the explanatory purposes of the observer or, conversely, make the detection time of a *sufficiently* precise modular description excessively high, even if the approximated algorithm is not, from a formal point of view, computationally hard. So, antimodularity, at least for what concerns detection of modularity requiring the algorithm to perform tasks which have been proved to be NP-hard, is a pragmatic but at the same time an *objective*, in a certain way unavoidable property of a system, deriving from computational properties which do not depend on contingent constraints or on a choice made by the observer, other than the choice of a required precision and of the basic description of the system.[27]

[27]The possibility of changing the basic description to avoid antimodularity seems in most cases precluded in real science, because each special science determines the basic description of its systems of interest: for example, molecular biology aims to describe a biological system in terms of molecules and their interactions. I think however this question should require further philosophical reflection.

6.3.1 Antimodularity Hinders Mechanistic and Functional Explanations

By examining its definition, it is easy to come to the conclusion that *antimodularity compels to single-level-only explanations*, excluding the possibility, highly desired for mechanistic and functional explanations, of an explanatorily valid *multi-level* description and the benefits it brings in terms of intelligibility and fine-tuning of the amount of conveyed information. Antimodularity, by its very definition, would limit mechanistic explanation to the level of description representing the most elementary parts of the system (the *basic description*), which is the most numerous level in terms of the number of parts and interactions among them, and the most complicated one. This fact hinders comprehension: for large enough systems, an explanation at this level is too complex to be understood by human beings, and understandability is a quality to be sought for in mechanistic explanation, at least according to the epistemic view on explanation which I, among others,[28] endorse.

With regard to functional analysis as described in Sect. 6.2.5, antimodularity would completely hinder the goal of obtaining a full multi-level hierarchy of functions, leaving us only the possibility of a two-level explanation: the highest level, the one of the explanandum itself (that is the global overall phenomenon to be explained) and, at the other end of the scale, the lowest level, the one of the most elementary functions (a "basic description" of the system). In this case, the observer, in order to give the system a functional explanation, would have to put each one of the elementary functions in relation to all the other elementary functions of the system and to the overall global function of the whole system which constitutes the explanandum. And this could, for complex enough systems, easily constitute a cognitively daunting task: the occurrence of antimodularity in a functional description means that such a description is graphically reduced to a large diagram composed of numerous low-level functional blocks, possibly interconnected in intricate ways. Such a diagram, when of a large enough size, could easily overcome human cognitive capabilities, yielding scarce to no comprehension of *how* the high-level functioning of the system is brought about. Differently from the case of mechanistic explanation (where one could always embrace the ontic view of mechanism, and content herself by considering the low-level mechanism *itself* as its known explanation) a low-level *purely functional* description which does not induce understanding is no explanation at all, because it lacks an ontic counterpart, and so we must conclude that *antimodularity dramatically hinders functional explanation*.

[28] See, again, Wright (2012).

6.4 Explanatory Emergence

Given that the lack of understanding due to the presence of antimodularity in a system can seemingly affect at least two kinds of explanation, the functional and the mechanistic one,[29] I propose a more general notion of *explanatory emergence*, understood as the property of systems[30] occurring when, *for absolute or pragmatic computational reasons*, they *resist understandable explanations*.[31] This is a more general property than antimodularity, comprising other possible effects of computational constraints on the explanation of complex systems. It does not apply only to automated computational tasks performed by computers, but also to possibly human operated tasks which bear on the intelligibility of explanations.

6.5 Discussion: Is It Likely to Encounter Antimodularity in Actual Science?

Antimodularity occurs when modularity detection turns out to be too computationally demanding to produce an explanatorily useful hierarchical description of the system in a feasible time. What is the likelihood that this circumstance can be encountered during actual, real-world scientific research?

It must be stressed that the computational complexity of modularity detection considered here concerns algorithms which do not employ any other informations about the system than those included in its basic description, the description of its elementary parts and of their relations: by adding ad hoc constraints, based on external contextual knowledge about the system, on how the elementary parts can be grouped into modules, the task can be highly simplified, and the corresponding ad hoc algorithms could end up being less computationally demanding than a general one. Actually, in many cases, this seems exactly what science does: it searches for empirical constraints to help us choose among the possible theories of the world. This increases the chance that scientific method can produce modular, intelligible descriptions of phenomena.

We must however wonder if new developments in science can shift the focus on systems of such size and complexity that even the known, empirically found constraints about them could end up being too few to allow for the successful

[29] Other types of explanations, such as deductive-nomological and computational explanation are affected too, as I intend to better highlight in a forthcoming work. Philippe Huneman's *topological explanation* (see Huneman 2015) is instead *enabled* by antimodularity, which in certain conditions is a topological property itself.

[30] Or *descriptions* of systems.

[31] I use the term "emergence" in a way akin to the conception, exposed in Ronald et al. (1999), of emergence as the appearance of something unexpected: in this case, the unexpected is the fact that a system is not explainable in an intelligible way.

completion of modularity detection on them. In the case of biological systems, we can actually be reasonably sure that they *are* modular, at least at certain levels of description.[32] Nevertheless, there can be significant biological systems, like for example interaction networks in the cell metabolism, which can end up being so huge, composed of so many parts and interactions between them, as to possibly produce effects of explanatory emergence due to the high computational cost of the algorithm for data mining or modularity detection in relation to the system's size. Antimodularity could thus possibly show up in *systems biology*, or *genomics*, or *connectomics*, where it is now normal to recur to computational methods for detecting modularity in complex systems.

These disciplines usually proceed nowadays, often aided by quick automated methods of laboratory analysis, by accumulating raw data about single protein-protein or protein-gene interactions, contributing this way to the construction of huge databases constituting very low-level representations of certain biological systems. Research then usually proceeds to automatically data-mine these databases in order to extract high-level structure from the low-level data. Research in these field is focusing (or could soon focus) on systems of such a complexity that the algorithm for their modularity detection could fail, due to the computational complexity of the required task.

There are signs of the actual occurrence of this condition: certain studies explicitly admit that the system, upon which modularity detection has been tested, had to be of limited size, because otherwise the algorithm would have taken too much time.

For example, Sales-Pardo et al. (2007) acknowledge:

> The computational cost of this step, the slowest one in our algorithm, limits network sizes to $\sim 10,000$ nodes. However, the cost can be reduced by using faster, but less accurate, methods for ordering the matrix, such as principal component analysis.[33]

Here we see the trade-off between accuracy of modularity detection and its cost. These, in a way, are *already* cases of antimodularity at work.

Again, in Orman et al. (2011), the authors admit:

> For time matters, it was possible to process networks with sizes 10,000 and 100,000, but not 500,000.[34]

The article is a survey and comparison of different community detection algorithms, so there is a paradox here: even assessments of the quality itself of community detection algorithms are hit by antimodularity, because of the difficulty in applying the modularity detection algorithms due to their computational cost.

[32] There are many arguments, empirical and theoretical, which favor this conclusion, as those by Herbert Simon (e.g. in Simon 1962) and by Stuart Kauffman (e.g in Kauffman 1993).

[33] Sales-Pardo et al. (2007, p. 15227).

[34] Orman et al. (2011, p. 273).

The situation seems more challenging for network motif detection, for which the actual approximate algorithms appear heavier: *Kavosh*,[35] which is one of the most accurate and fast algorithms, would employ 17 million seconds (more than six months) to search for 8-nodes motifs in the *Saccharomyces cerevisiae*'s transcriptional network, a network composed of only 688 nodes and 1079 edges. The same algorithm would employ 300 million seconds (almost 10 years) to search for 9-nodes motifs on the same network, or more than 200 years for 10-nodes motifs.[36] Moreover, certain approximate algorithms which make use of probabilistic approaches by sampling the original graph in order to reduce the complexity of the task[37] run the risk of failing to recognize the occurrence of certain motifs in the parts which do not get sampled.

When it happens that the complexity of the task associated with the size of the input data compels to resort to too approximate methods for modularity detection, and consequently the hierarchical modular description obtained is too simplified to be explanatorily valid in relation to the explanatory goals of the researcher, this would constitute an occurrence of antimodularity in actual science. Excessive computational complexity could occur also during the algorithmic data mining aimed to extract different kinds of structure, other than modularity, from the raw data, and when this computational cost compels to resort to a too simplified description, which appears then inadequate for the explanatory goals of the researcher, this would constitute a case of explanatory emergence.

Some concern about antimodularity should probably affect systems biology. This discipline has always aimed at giving a high-level (or multilevel) mechanistic explanation of whole biological systems, and this is perfectly in line with the recourse to automated methods for detecting the system's functional modular high-level organization. But mechanistic explanation as such, and thus also *high-level* mechanistic explanation, must be based on a sufficiently *exact* functional and structural decomposition of the system, otherwise the obtained high-level explanation is not valid and consequently it is not an explanation at all. This condition requires a further, at least necessary (even if, probably, not sufficient) condition: the obtainment of high-level modular descriptions of the system under observation which are *accurate enough*. But the problematic features of the most used quality metrics of modularity, combined with the proved computational intractability of the optimization of certain among them, puts the perspective of obtaining sufficiently accurate high-level functional modular characterizations of a system at risk of being frustrated, if the system is composed by a large enough number of elementary parts. This is due to the fact that computational hardness in many cases makes the approximate algorithms for optimizing the quality measure subject to a trade-off between speed and precision.

[35] Introduced in Kashani et al. (2009).

[36] As reported in Wong et al. (2012). See p. 9, Table 4 and p. 12, Table 5 in that paper.

[37] See, again Wong et al. (2012).

In general, when a functional hierarchical representation of the system is sought for, a high precision of this representation is needed, especially if the system is complex and highly non-linear, a circumstance which often occurs in natural systems. Antimodularity would produce inexact representations of the system. In the case of complex non-linear systems, an error resulting in a non accurate partition of the system, maybe mistaken by only a few communities, or one that neglects the occurrence of some motifs, as is typical of some approximate algorithms, could easily produce a completely invalid (from a dynamical and functional point of view) description of the system, and as such an explanatorily useless description.

That said, it seems plausible that we will have difficulty in coming upon reports of actual cases of antimodularity in the scientific literature, and this for a simple reason: as highlighted in many bibliographical studies,[38] it is nowadays very unlikely, if not impossible, that negative results get ever published. And, antimodularity would compel the authors to admit that they have *failed* in finding a valid high-level explanation of a complex system due to the failure of algorithmic methods for extraction of a high level structure from the data. It seems unlikely that such an admission could ever be published, or even *proposed* for publication by the researcher.

6.6 Conclusions

Due to its nature of being dependent on the unavailability of feasible good approximations of certain computationally intractable tasks, antimodularity is a subtle circumstance: avoiding it depends on our capacity to devise more and more intelligent approximate methods to circumvent, at least partially, the exact task's computational intractability.

An external phenomenon, however, seems capable to render the occurrence of antimodularity increasingly likely: the more we manage to algorithmically extract sufficiently accurate modular hierarchical descriptions of complex systems, and to produce this way acceptably valid high-level functional and mechanistic explanations of their behavior, the more we are pushed toward attacking larger and more complex systems with these same good approximate algorithmic methods: this is a practical positive feedback loop in the application of science, which self-increases. But like in an arms race, at some cycle or another, the size of the input data could overcome the ability of the best current algorithms to extract modular structure from them, and so antimodularity could occur.

All in all, due to the improbability that admissions of this kind of failure get declared in the literature, antimodularity could for a long time simply *lurk* behind scientific research conducted analyzing big data. Antimodularity will probably show up clearly only in the case its presence became, in time, so widespread to begin

[38] See, for example, Smaldino and McElreath (2016).

hindering entire branches of science. This may occur very soon or very far in the future, maybe never. But I think that, as a matter pertaining to philosophy of science, this eventuality could be well worth a preliminary reflection.

References

Barber, Michael J. 2007. Modularity and community detection in bipartite networks. *Physical Review E* 76. https://doi.org/10.1103/PhysRevE.76.066102

Bechtel, William, and Adele Abrahamsen. 2005. Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441. Mechanisms in Biology. https://doi.org/10.1016/j.shpsc.2005.03.010

Bechtel, William, and Robert C. Richardson. 1993. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008: P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Brandes, U., D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. 2008. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering* 20: 172–188. https://doi.org/10.1109/TKDE.2007.190689

Chakraborty, Tanmoy, Ayushi Dalmia, Animesh Mukherjee, and Niloy Ganguly. 2016. Metrics for community analysis: A survey. *arXiv:1604.03512 [physics]*.

Cummins, Robert. 1975. Functional analysis. *The Journal of Philosophy* 72: 741–765. https://doi.org/10.2307/2024640

Danon, Leon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* 2005: P09008. https://doi.org/10.1088/1742-5468/2005/09/P09008

Garey, Michael R., and D.S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of Np-Completeness*. New York: W. H. Freeman.

Girvan, Michelle, and Mark E.J. Newman. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99: 7821–7826. https://doi.org/10.1073/pnas.122653799

Good, Benjamin H., Yves-Alexandre de Montjoye, and Aaron Clauset. 2010. Performance of modularity maximization in practical contexts. *Physical Review E* 81. https://doi.org/10.1103/PhysRevE.81.046106

Hempel, Carl G., and Paul Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of Science* 15: 135–175. https://doi.org/10.1086/286983

Huneman, Philippe. 2015. Diversifying the picture of explanations in biological sciences: Ways of combining topology with mechanisms. *Synthese* 1–32. https://doi.org/10.1007/s11229-015-0808-z

Kashani, Zahra, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. 2009. Kavosh: A new algorithm for finding network motifs. *BMC Bioinformatics* 10: 318. https://doi.org/10.1186/1471-2105-10-318

Kauffman, Stuart A. 1993. *The origins of order: Self-organization and selection in evolution*. New York: Oxford University Press.

Kreinovich, Vladik, and Max Shpak. 2006. Aggregability is NP-hard. *ACM SIGACT News* 37: 97–104. https://doi.org/10.1145/1165555.1165556

Kreinovich, Vladik, and Max Shpak. 2008. Computational aspects of aggregation in biological systems. In *Applications of Computational Intelligence in Biology*. Studies in Computational Intelligence, vol. 122, ed. Tomasz G. Smolinski Dr, Professor Mariofanna G. Milanova, and Professor Aboul-Ella Hassanien, 281–305. Berlin/Heidelberg: Springer.

Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. Thinking about mechanisms. *Philosophy of Science* 67:1–25. https://doi.org/10.1086/392759

Miyauchi, Atsushi, and Noriyoshi Sukegawa. 2015. Maximizing Barber's bipartite modularity is also hard. *Optimization Letters* 9:897–913. https://doi.org/10.1007/s11590-014-0818-7

Newman, Mark E.J. 2003. The structure and function of complex networks. *SIAM Review* 45: 167–256. https://doi.org/10.1137/S003614450342480

Newman, Mark E.J., and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69: 026113. https://doi.org/10.1103/PhysRevE.69.026113

Orman, Günce Keziban, and Vincent Labatut. 2009. A comparison of community detection algorithms on artificial networks. In *Discovery Science*. Lecture Notes in Computer Science, vol. 5808, ed. João Gama, Vítor Santos Costa, Alípio Mário Jorge, and Pavel B. Brazdil, 242–256. Berlin/Heidelberg: Springer.

Orman, Günce Keziban, Vincent Labatut, and Hocine Cherifi. 2011. Qualitative comparison of community detection algorithms. In *Digital Information and Communication Technology and Its Applications*. Communications in Computer and Information Science, vol. 167, ed. Hocine Cherifi, Jasni Mohamad Zain, and Eyas El-Qawasmeh, 265–279. Berlin/Heidelberg, Springer.

Papadopoulos, Symeon, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. 2011. Community detection in social media. *Data Mining and Knowledge Discovery* 24: 515–554. https://doi.org/10.1007/s10618-011-0224-z

Plantié, Michel, and Michel Crampes. 2013. Survey on social community detection. In *Social Media Retrieval*, ed. Naeem Ramzan, Roelof van Zwol, Jong-Seok Lee, Kai Clüver, and Xian-Sheng Hua, 65–85. London: Springer.

Ronald, E., M. Sipper, and M. Capcarrère. 1999. Design, observation, surprise! A test of emergence. *Artificial Life* 5: 225–239. https://doi.org/10.1162/106454699568755

Sales-Pardo, Marta, Roger Guimerà, André A. Moreira, and Luís A. Nunes Amaral. 2007. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences* 104: 15224–15229. https://doi.org/10.1073/pnas.0703740104

Shen-Orr, Shai, Ron Milo, Shmoolik Mangan, and Uri Alon. 2002. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics* 31: 64–68. https://doi.org/10.1038/ng881

Simon, Herbert A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* 106 (6): 467–482.

Simon, Herbert A., and Albert Ando. 1961. Aggregation of variables in dynamic systems. *Econometrica* 29: 111–138. https://doi.org/10.2307/1909285

Smaldino, Paul E., and Richard McElreath. 2016. The natural selection of bad science. *arXiv:1605.09511 [physics, stat]*.

Winker, Peter. 1992. *Some Notes on the Computational Complexity of Optimal Aggregation*. 184. Diskussionsbeitrãge: Serie II, Sonderforschungsbereich 178 "Internationalisierung der Wirtschaft", Universität Konstanz.

Wong, E., B. Baur, S. Quader, and C.-H. Huang. 2012. Biological network motif detection: Principles and practice. *Briefings in Bioinformatics* 13: 202–215. https://doi.org/10.1093/bib/bbr033

Wright, Cory D. 2012. Mechanistic explanation without the ontic conception. *European Journal for Philosophy of Science* 2: 375–394. https://doi.org/10.1007/s13194-012-0048-8

Yang, Bo, Dayou Liu, and Jiming Liu. 2010. Discovering communities from social networks: Methodologies and applications. In *Handbook of Social Network Technologies and Applications*, ed. Borko Furht, 331–346. Boston: Springer.

Chapter 7
A Software-Inspired Constructive View of Nature

Russ Abbott

Abstract In their review article on "Scientific Reduction" Van Riel and Van Gulick (Scientific reduction. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Spring 2016 edition). Stanford University, Stanford, 2016) write,

> Saying that x reduces to y typically implies that x is *nothing more than* y or *nothing over and above* y.

The y to which an x reduces consists most often of x's components. But virtually nothing can be reduced if to be "nothing more than" or "nothing over and above" its components means to have no properties other than those of its components, individually or aggregated. An atom has properties other than those of its quarks and electrons. A protein, a biological cell, and a hurricane—not to mention such man-made entities as houses, mobile phones, and automobiles—all have properties over and above their components. The properties of most entities depend on *both* those of the entity's components *and* on how those components are put together. (That would seem obvious, but perhaps it's not.)

One of the defining characteristics of what might be referred to as the creative disciplines—computer science, engineering, the creative arts, etc.—is a focus on understanding and using the effects of putting things together. They ask what new (and in human terms interesting and useful) properties can be realized by putting things together in new ways. Using software as an example I explore software construction, and I ask what, if anything, one gains by thinking of it reductively.

Reduction as nothing-more-than-ism tends to blind one to nature's constructive aspects. I discuss nature's tools for creating new phenomena, including negative interactive energy, means for creating and tapping stores of usable energy, autopoiesis, and biological evolution.

R. Abbott (✉)
Department of Computer Science, California State University, Los Angeles, CA, USA
e-mail: RAbbott@calstatela.edu

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_7

Keywords Reductive explanation · Creative construction · Negative interaction energy · Evolution · Causation

7.1 Is There More to Nature Than Physics?

Four physics Nobel laureates give conflicting answers.[1] In his gentle way, Einstein (1918) argued that physics explains everything.

> The painter, the poet, the speculative philosopher, and the natural scientist each in his own fashion, tries to make for himself a simplified and intelligible picture of the world. What place does the theoretical physicist's picture of the world occupy among these?
>
> The physicist contents himself with describing the simplest events that can be brought within the domain of experience. But what can be the attraction of getting to know such a tiny section of nature while leaving everything subtler and more complex shyly and timidly alone? Does the product of such a modest effort deserve to be called by the proud name of a theory of the universe?
>
> In my belief the name is justified; for the general laws on which the structure of theoretical physics is based claim to be valid for any natural phenomenon whatsoever. With them, it ought to be possible to arrive by pure deduction at the theory of every natural process, including life.

Weinberg (2003) agreed.

> One can illustrate the reductionist world view by imagining all the principles of science as dots on a huge chart, with arrows flowing into each principle from the principles by which it is explained. History shows that these arrows do not form disconnected clumps, representing independent realms of science; and they do not wander aimlessly. They are all connected, and if followed backward they all seem to branch outward from a common source, an ultimate law of nature. Thus the reductionist regards the general theories governing air and water and radiation as being at a deeper level than theories about cold fronts or thunderstorms: the latter can in principle be understood as mathematical consequences of the former. Similarly, apart from historical accidents that by definition cannot be explained, the rules governing phenomena like mind and life have evolved to what they are entirely because of the principles of macroscopic physics and chemistry, which in turn are what they are entirely because of the principles of the standard model of elementary particles.
>
> The reductionist program of physics is the search for the common source of all explanations. We hope that in the future we will have achieved an understanding of all the regularities that we see in nature, based on a few simple principles, laws of nature, from which all other regularities can be deduced.

Schrödinger (1944) was not convinced.

> Living matter, while not eluding the 'laws of physics' is likely to involve "other laws," which will form just as integral a part of its science.

Anderson (1972) agreed with Schrödinger.

[1] Extracts are slightly paraphrased throughout the paper.

> The workings of all the animate and inanimate matter of which we have any detailed knowledge are controlled reductively by the fundamental laws of physics, which I fully accept. Even so, the behavior of large and complex aggregates of elementary particles is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear.

This paper explores the nature and status of Schrödinger's "other laws" and Anderson's "new properties." Section 7.2 reviews the current philosophical status of explanatory reductionism[2] and concludes that it cannot serve as the primary scientific paradigm. Section 7.3 discusses reverse engineering as a possible replacement. Since reverse engineering is a form of analysis, Sect. 7.3 also discusses Anderson's asymmetrical contrast between analysis and synthesis.

Sections 7.4 and 7.5 are the heart of the paper. Section 7.4 uses an example from Computer Science to introduce what I'm calling the constructive perspective. A constructive perspective requires means to construct new entities. Section 7.5 discusses what I consider to be nature's primary tools of construction.

Section 7.6 examines effective field theory as an alternative approach to scientific explanation. Section 7.7 discusses an example of a construction that depends on mathematical truth as well as physical properties. Section 7.8 offers some brief concluding remarks.

7.2 Is Reductionism Dead? What Philosophers Say

Fodor has long doubted (1974 and 1997) that fundamental physics can explain higher level regularities.

> **Molto *Mysterioso*.** Damn near everything we know about the world suggests that unimaginably complicated to-ings and fro-ings of bits and pieces at the extreme *micro*-level manage somehow to converge on stable *macro*-level properties. The 'somehow' really is entirely mysterious. How can macro-level stabilities supervene on a buzzing, blooming confusion of micro-level interactions? Why is there anything except physics? Well, I admit that I don't know why. I don't even know how to think about why. I expect to figure out why there is anything except physics the day before I figure out why there is anything at all. –Fodor (1997)

Loewer (2009) characterized what he called a lightweight version of non-reductive physicalism.

> The special sciences contain vocabulary/concepts that are conceptually independent of the concepts and vocabulary of physics. A biologist may have evidence that a biological generalization is lawful (think of the Mendelian laws) without having any idea how this regularity is rendered lawful or implemented by fundamental laws of physics.

[2]Brigandt and Love (2015) distinguish two primary categories of reduction: "*theory reduction*, the claim that a higher level theory can be logically deduced from a lower level theory, and *explanatory reduction*, the claim that representations of higher level features can be explained by representations of lower level features, typically by decomposing a higher level system into parts." I focus primarily on explanatory reduction.

> The nomological structure of the world is completely specifiable by fundamental physics. The special sciences characterize aspects of that structure that are especially salient to us and amenable to scientific investigation in languages other than the language of physics.

The remainder of this section reviews, briefly but broadly, the current state of philosophical reductionism. I'll refer primarily to Van Riel[3] and Van Gulick (2016) and Brigandt and Love (2015).

Van Riel and Van Gulick quote what they refer to as Smart's (1959) tentative but influential formulation of what is required for a reduction.

> An entity *x* reduces to an entity *y* only if *x* does not exist 'over and above' *y*.

It's not clear—at least to me—what *over and above* means in this context. Below we discuss a hydrogen atom. It consists of a proton and an electron held together by electromagnetic attraction and obeying certain rules of quantum mechanics. Is such an atom *nothing over and above* its constituents?

I would say that a hydrogen atom is more—actually less (see the discussion below for why)—than its component particles. Once the question is raised it's difficult to think of anything other than simple aggregations that fails to have properties over and above the aggregated properties of its components. So it's not clear how to make sense of Smart's suggested criterion.

Brigandt and Love (2015) focus on biology, but their discussion generalizes. The following outlines the reductionist[4] case.

- Each biological system (e.g., each organism) is constituted by nothing but molecules and their interactions.[5]
- Biological properties supervene on physical properties.
- Each particular biological process is metaphysically identical to some particular physico-chemical process.

Like Anderson and Schrödinger virtually no philosophers or scientists argue for material phenomena independent of physics.[6] But compatibility with physics is not

[3]Gerd Van Riel seems to spell his name at different times with an upper or lower case "V." His current affiliation, KU Leuven, (http://www.kuleuven.be/wieiswie/en/person/u0019425, retrieved Sep 3, 2016) uses upper case.

[4]Brigandt and Love also discuss *methodological reduction* ("biological systems are most fruitfully investigated at the lowest possible generally biochemical level)" and *epistemic reduction* ("knowledge about one scientific domain can be reduced to knowledge about a more fundamental level").

[5]This sounds uncontroversial, but the exchange of matter and energy between biological organisms and the environment raises questions. When does an oxygen or food molecule become part of an organism? When does waste matter become not part of an organism? Are organ transplants or component implants such as pacemakers, corneal replacements, tooth implants, or dental fillings part of an organism? What about disease agents and toxins? What about an organism's biome and virome? Does a photon that conveys visual information become part of an organism—and if so when? Given these considerations it seems a much more complex issue to decide what constitutes a biological system than this innocent-sounding statement implies.

[6]Physics being what it is, the discovery of any such phenomena would inevitably lead to "new" physics anyway.

the point. The problem is that many phenomena challenge a standard mereological perspective. For example, MacLeod and Nersessian (2015) write that "one of the central claims of systems biology is that properties and biological functions of components are dependent on their participation within systems."

(a) *Phenomena in context.* Folded functional proteins consist of linked chains of amino acids. Yet other already-folded proteins must be present to assist in the folding process. The amino acid components alone are insufficient causally, even if they are sufficient constitutionally.
(b) *Dynamic phenomena*. Higher level (sometimes called "self-organizing") regularities such as flocking and bacterial quorum sensing result from behavioral interactions among system components rather than from the characteristics of the components individually.

Brigand and Love summarize the consequences of these and other considerations as follows.

> It is more difficult to conceptualize a single, adequate conception of reduction that will do justice to the diversity of phenomena and reasoning practices in the life sciences. Consequently, some philosophers suggest that we should move beyond reductionism entirely.

I agree. One may characterize the relationship between certain types or certain theories as reductive, but evidence is dwindling that reductionism *per se* will serve as a unifying mechanism for science.

7.3 If Science Isn't Primarily Reduction, What Is It?

Most scientists are interested in understanding nature. Weinberg talked about chains of explanation. Dawkins (1986) expressed a similar view a decade and a half earlier: "Reductionism is simply an honest desire to understand how things work."

The great work of science involves two broadly defined tasks.

Task 1. To observe and characterize categories of phenomena that have identifiable regularities.
Task 2. To explain how those regularities come to be.

Task 1 often requires the development of new models and representations for describing the regularities. The equations of quantum theory describe observed phenomena and their regularities, but they do not explain how those phenomena are brought about. Feynman (1964) pointed out that Aztec astronomers had quite accurate calendar-based formulas for predicting eclipses. But they had no conceptual models to explain what brings eclipses about.

Task 2 involves determining how some regularity might come about. For the Aztec eclipse formulas, Task 2 would involve building a model that includes such

elements as the moon orbiting the earth and the earth orbiting the sun and showing how that model produces the eclipses predicted by their formulas.

Engineers and computer scientists often face Task 2-like challenges. They are confronted with an operational system whose internal workings are hidden—either deliberately for commercial reasons or because the relevant documentation was lost or never produced. They are asked to explain and perhaps recreate the system's functionality. This task is known as reverse engineering.

This is also the work of science. Scientists are confronted with observed phenomena and want to know how nature managed to bring them about. To answer the question posed in the section header, *one of the grand goals of science is to reverse engineer nature.*

Isn't this just reductionism by another name? No. Recall the two example from systems biology. In both cases careful reverse engineering enabled scientists to identify the responsible mechanisms. But those mechanisms were not reductive. System behavior depended on the structure and organization of the system itself and could not be explained strictly in terms of the aggregated system components.

7.3.1 Is Synthesis More Difficult Than Analysis?

Anderson (1972) contrasted scientific analysis with what he considered its opposite.

> The ability to reduce everything to simple fundamental laws, i.e., *the reductionist hypothesis*, does not imply *the constructionist hypothesis*, namely, the ability to start with those laws and reconstruct the universe. The relationship between a system and its parts is intellectually a one-way street. Scientific analysis may be not only possible but fruitful. Synthesis is expected to be all but impossible.

But other than to say that "More is Different" Anderson didn't explain in any detail why he thought synthesis would be so difficult. Consider these two tasks.

- *The construction (implementation[7]) problem*: find a way to use elements and processes from a collection of resources to produce a desired phenomenon.
- *The analysis (realization) problem*: given a known phenomenon, determine how elements and processes from a collection of resources could have produced it.[8]

Expressed in those terms the two problems are quite similar—but not identical. For example, one doesn't necessarily know whether implementation of a specified phenomenon is possible; one knows for sure—since one observes it—that the

[7] As I'm using the terms, to *implement* is to create something that has certain pre-specified properties. Engineers *implement* systems with required functionality. To *realize*—as in the heart *realizes* a pumping capability—is for something to come into existence that happens to have some properties. Nature *realizes* various functionalities without having those functionalities as teleological goals. See Abbott (2016a) for an expanded discussion.

[8] Wilson (1998) called this *recomposition*: to show that components can be reassembled to "capture the key properties of the entire ensembles."

phenomena to be analyzed exist. Yet in terms of intellectual difficulty the two problems are roughly comparable. Why did Anderson consider synthesis harder?

The challenge of synthesis lies in its open-endedness. Imagine that instead of being asked to solve a particular implementation problem, one were asked to come up with, say, the full range of ways in which living things could persist in the world. Considering life's enormous variety one would understandably find that task all but impossible. Not only has life found an extraordinary number of ways to sustain itself, but new strategies are continually being evolved. Many bacteria have solved the problem of living in a world of anti-bacterial agents. Some can live with all known anti-bacterial agents.

Furthermore, one can imagine solving the general analysis problem—the dream of Einstein and Weinberg. In contrast, synthesis never ends. The range of new possibilities continually expands.

Reverse engineering helps with the analysis problem but is of limited help with synthesis.

7.4 The Constructive Perspective of Computer Science

This section introduces the constructive perspective of computer science and contrasts it with the analytic perspective of reductionism.

Consider the following three lines of software pseudocode.

```
temp ← x;
x ← y;
y ← temp;
```

After these three lines are executed, the values initially stored at `x` and `y` will have been exchanged: `y` holds what `x` originally held and *vice versa*.

The code itself is transparent. No further explanation is needed once one understands that

(a) `x`, `y`, and `temp` name places where values may be stored;
(b) the ← operation copies the value stored at the location named to its right and stores that copy at the location named to its left; and
(c) the three lines are performed in the order written,

Suppose we took these three lines, bundled them into a unit with `x` and `y` as parameters—perhaps called an *exchange* operation—and wrapped it opaquely so that the interior is not visible. We could then use it to perform exchanges for us. It's phenomenology is straightforward: it exchanges `x` and `y`.

Suppose that years later an observer, wondering how it worked, took it apart and discovered the three lines. Should she consider those lines to be a *reduction* of this *exchange* operation?

Recall that Van Riel and Van Gulick adopt Smart's criterion for reduction.

> Saying that *x* reduces to *y* typically implies that *x* is *nothing more than y* or *nothing over and above y*.

I argue that it's not correct to say that the *exchange* operation as a unit is *nothing more than* or *nothing over and above* the aggregation of the three lines.

- The unit itself is a stable entity, not just three lines that happen to be physically adjacent.

The unit has unit-level properties, including at least the following.

- The lines are to be performed in an indicated order.
- If the same name, e.g., x, appears multiple times, it refers to the same location each time.

So the exchange unit is more than just three separate lines. It is those lines *along with* the properties just listed. How is such a unit created? One uses what is known as a *constructor*. Constructors enable one to build new things by combining existing things in particular ways.

Constructors came into their own with object oriented programming languages. An object is an organized collections of pre-existing elements that are joined together in a particular way.

Software objects are not metaphorical. Once created an object may be treated as an entity. Like entities, an object may be named and referred to as a whole. An object may be stored as the reference of a variable. An object may be passed to a function as an argument.

Although common in computer science, as far as I can tell the notion of a constructor appears rarely if at all in philosophy. A search[9] of the *Stanford Encyclopedia of Philosophy* found *constructor* in only 6 articles, of which only Dybjer and Palmgren (2016) use *constructor* in the computer science sense. Yet as we'll see, constructors, i.e., mechanisms that builds new entities, play an important role in nature.

7.4.1 Is Software a Reduction of a Software-Defined World?

There is even a larger question. Is reduction a useful way of thinking about the relationship between a software system and the functionality it implements.

Software has two properties that makes it an interesting case about which to ask this question. (a) One can analyze software into components, as in the example above. (b) Software primitives, i.e., machine instructions, are well understood.

[9] http://plato.stanford.edu/search/search?query=constructor, conducted Sep 8, 2016.

Let's consider a fairly complex software system that tracks student enrollment. The system's functionality would be described in terms of students, courses, (academic) departments, degree programs, course sections, instructors, classrooms, etc., the sorts of things one would expect when talking about students enrolling in courses. Interactions among instances of these types include such activities as a student enrolling in a class, an instructor being assigned a class to teach, an instructor assigning a grade to a student for a course, etc.

Is such a software system a reduction of the world of academic entities to the world of programming language entities? In the language of Nagel-style reduction (1970): the programming language defines the laws and properties of programming language entities, and the system specification defines the laws and properties of the academic entities. Software bridges the gap and serves as bridge laws.

Software bridges such gaps in what is often painfully complete and tedious detail. Many software systems involve millions of lines of code. Would Nagel have considered a reduction that included bridge laws consisting of millions of lines of often complex logic as fitting his notion of what a reduction should consist of? Of course we don't know. But he may have agreed with Anderson that more is different.

Whether or not Nagel would have considered software bridge laws a valid part of a reduction, what do we gain by thinking of it this way? We already know that the model of academic entities is implemented by software. What do we gain by applying the term *reduction* to the implementation? Doing so doesn't help us understand how the system works. We understand how the system works by looking at how the software *constructs* the world of academic entities, not by thinking of the software as a mapping between the laws of programing language entities and academic entities.

Furthermore, we generally design complex software system in terms of libraries, modules, frameworks, and other components—not directly in terms of programming language entities. It may be possible to describe the design of a software system in terms of its lowest level elements, i.e., bits and machine instructions, but the result would be so complex as to be useless for understanding how it works.

The same holds for science generally. Fodor was right. Higher level regularities result from the interaction of entities that obey the laws of physics. But to trace the relationship between the two in a way that makes the resulting analysis useful would involve intermediate level entities in much the same way that tracing how software implements a complex system involves intermediate-level components.

What benefit is gained by using the term *reduction* to refer to a relationship between (depending on the example) (a) entities of a programming language or elementary particles of physics and (b) higher level entities? I would say that the term *reduction* adds little if anything in either case.

7.4.2 Perhaps Nothing Is Gained, But Is Anything Lost?

Higher level regularities result from particular organizations of lower level phenomena. These organizations are said to implement or realize[10] the higher level phenomena. The terms *implementation* and *realization* imply constructive activities. The term *reduction* seems to dismiss the constructive aspect of the relationship—and in doing so to leave out something fundamentally important. To construct something implies—at least in this context—that the resulting construction holds together as an entity with certain properties.

As suggested above, the challenge posed by synthesis arises from nature's open-endedness. New possibilities continually present themselves. Many are realized. Because a constructive perspective focuses on building new things, it keeps us aware of nature's open-endedness. A reductive perspective focuses on existing things and how they may be composed of simpler things. In doing so it tends to ignore nature's constructiveness and open-endedness.

When describing what he called generative grammars, Chomsky (1966) referred to the

> creativity of language, the ability of speakers of a language to produce and interpret sentences similar to sentences that have been heard before only in that they were generated by the rules of the same grammar.

The generative nature of grammars has turned out to be of fundamental importance. Similarly, at least in my view, many of the important questions about nature concern the mechanisms that enable the continual production of new, stable, and often increasingly complex and sophisticated phenomena.

Software development is a constructive activity. When writing new software one builds something new from things that already exist. A purely reductive view misses that perspective.

One could no more understand the functioning of a software system by looking at its components in isolation than one could understand the harmonic and melodic effect of a piece of music by looking at just the individual notes. Nor does it make sense to say that a piece of music is *nothing more than* or *nothing over and above* its components. In both cases, the organization of the components is crucial to how the system or composition produces the phenomena it does.

Recall Fodor's plaintive cry.

> ***Molto* Mysterioso.** Unimaginably complicated to-ings and fro-ings at the extreme *micro*-level manage somehow to converge on stable *macro*-level properties. The 'somehow' really is entirely mysterious.

One might find large software systems similarly mysterious. Computer games, software that models cell biology, theorem proving software, software that beats the best humans at Go and poker. How can such macro-level functionalities supervene

[10] Recall our earlier discussion of *implementation* and *realization*.

on a buzzing, blooming confusion of micro-level bits and instructions? Yet they do. It's not mysterious. It only seems mysterious if one ignores the multiple levels of constructive capabilities software developers have invented. One loses that constructive perspective when thinking in strictly reductive terms. The primitives are known. What matters is how to put them together.[11]

7.5 How Nature Works as a Constructive Open-Ended System

What we haven't talked about is what enables one to build new things from existing pieces. In software it's easy. The software framework provides the tools. Programming languages have constructors and other compositional operations built into them.

Similarly in music, one simply writes notes on a piece of paper—or now on a computer screen. Once written, the paper or computer holds them in a fixed relationship. Software and music inhabit synthetic worlds that hold objects together essentially by definition. One can simply posit a software object or write down a musical chord, and it exists. See Abbott (2010).

But how does nature build new things? I will discuss four features that underlie nature's constructive and generative capabilities: negative interaction energy, energy management mechanisms, autopoiesis, and biological evolution.

7.5.1 Negative Interaction Energy

How does the world beyond elementary particles come to be as it is? We know, for example, that a hydrogen atom consists of a proton and an electron. Why do they stay together as an entity?

The answer has to do with what physicists call interaction energy.[12] Interaction energy can be positive or negative. It corresponds to what one may intuitively think of as a force pushing things apart or pulling things together. The repulsive force between objects with the same electric charge and the attractive force between objects with different electric charges is best understood as positive or negative interaction energy respectively. Similarly for gravity. Its attractive force is a consequence of negative interaction energy. Quoting Strassler (2015),

[11] Physics is still investigating the primitives. But given the primitives as currently understood, the point still holds.

[12] Although Strassler (2015) coined the term, the phenomenon (under various names) is fundamental to physics.

> The possibility that interaction energy can be negative is the single most important fact that allows for all of the structure in the universe, from atomic nuclei to human bodies to galaxies.

How does this relate to the hydrogen atom? When the magnitude of the negative interaction energy—in the form of electromagnetic attraction—between a proton and an electron overcomes the particles' kinetic energies—and other situation-specific influences—the two become a hydrogen atom. In other words *the laws of physics provide means for entities to come together to form stable compound entities*. Negative interaction energy is the glue that enables nature to build stable new entities.

7.5.1.1 General Evolution

Negative interaction energy and the compounds it produces give rise to what might be considered a more general form of evolution: new compound entities are created as combinations of existing entities. In many ways this resembles ecological succession: as a consequence of being populated by certain entity types, an environment becomes more suitable for other entity types. History-of-the-universe infographics[13] illustrate this phenomenon. The universe evolved

- from a plasma soup at the big bang
- to a baryon soup spiced with photons, neutrinos, and electrons
- to collections of protons, neutrons, atomic nuclei, atoms, and molecules
- to stars, planets, and galaxies, which gave birth to
  - heavier atomic elements via nuclear reactions in the interior of stars and
  - molecular combinations and other mixtures via chemical and mechanical interactions in environments like the earth and in interstellar dust clouds (e.g., McGuire et al. 2016).

This process has produced extraordinary results. Besides atomic elements, molecules, planetary systems, black holes, galaxies, etc., it has also produced such cosmic features as quasars and pulsars and planet-level features as weather (which moves massive amounts of materials across a planet and which also produces storms, which in turn create canyons, etc.), geological activity (such as – volcanoes, hot springs, earthquakes, continental drift, hydrothermal vents), solar flares, planetary rings, etc. General, i.e., non-biological, evolution can be a powerful and creative process. It reflects one aspect of nature's ability to generate new phenomena. Ultimately it relies on negative interaction energy.

[13] See example images from Lawrence Berkeley National Lab: http://particleadventure.org/history-universe.html

7.5.1.2 Less-Mass Emergence

Are entities built using negative interaction energy more than the aggregation of their components? Actually, they are less.

The mass of a hydrogen atom is the sum of the masses of its proton and electron *plus* the mass equivalent of the interaction energy that holds it together. Since the interaction energy in this case is negative, the mass of a hydrogen atom is *less than* the sum of the masses of its proton and electron considered separately. To separate a hydrogen atom into its constituents requires the *addition* of enough mass (in the form of energy) to make up for the negative interaction energy. A hydrogen atom exists in an energy well. To lift it out—to pull the electron off the nucleus—requires energy.

More generally, compounds bound together by negative interaction energy typically have different properties (or different values of a given property) than their components considered separately. Whatever one means by *entity*[14] (or entity type or kind) it would seem perverse to argue that such compounds should not count as distinct entities (or entity types or kinds).[15]

7.5.2 Energy Accumulation, Storage, and Release

Classical physics defines usable energy as energy available to do work.[16] An alternative version of the Second Law of Thermodynamics is that usable energy always decreases overall. Here *decreases* means something like drains off into the environment in such a way that it is no longer available to do work. Since energy is conserved, the drained off energy is not lost. It just becomes indistinguishable from ambient energy and cannot be put to use.

Stores of usable energy can be created—for a price. A naturally occurring example is evaporation of water vapor, whereby water moves from ground level to high in the atmosphere where it has usable energy in the form of gravitational potential energy. (It is "used" during rain and snow storms. It carves canyons, and

[14] I was unable to find a relevant philosophical analysis of the term *entity*. Bricker (2014), for example, discusses ontological commitment but not an analysis of the grounds that justify ontological commitment.

[15] Such phenomena account for many of the things traditionally considered emergent. Abbott (2010) called this *static emergence*. It is *static* because the results are statically stable entities. This mechanism is not responsible for aggregations or software objects, which are not bound together by negative interaction energy.

[16] Work is defined as force applied over distance. I don't know whether there is a general term for energy available for doing work. *Thermodynamic free energy* may not cover all cases, e.g., the kinetic energy of a bowling ball rolling down an alley or energy transferred from a planet to a satellite during a gravitational slingshot maneuver. Even isolated orbiting masses radiate energy—reflecting the consumption of usable energy. See Koberlein (2016).

we exploit it as hydropower.) The amount of solar energy needed to move water vapor exceeds the amount of usable energy stored. As in energy transformations generally, conversion efficiency is less than 100%.

Nature has discovered many ways to create, store, and release usable energy. These include gravitational compression and then nuclear reactions in stars, geological motion (which stores energy as stress, which is released as earthquakes), and the just mentioned meteorological activity. Biological mechanisms for creating, storing, and releasing usable energy include (a) photosynthesis and (b) eating other organisms and using their carbohydrates, proteins, and fat as energy sources. Biological organisms also store energy chemically as ATP and as gradients across membranes. These all reduce the total amount of usable energy in the universe, but locally they capture and store usable energy.

Why does this matter? The most interesting natural phenomena occur when stored energy is "spent," and especially when it is released in a way that realizes various functionalities—the closest nature comes to teleological phenomena. This is the subject of the next section.

7.5.3 Switches and Autonomous Causality

This section focuses on mechanisms for managing the use of accumulated energy. I'll talk first about switches, an indirect way to manage energy flows, and then autonomous causality, a more indirect and more sophisticated way to manage energy flows. See Abbott 2016b for further discussion.

7.5.3.1 Causality

Causality provides a useful framework of the discussion of energy flows. Following are two of the most widely used characterizations of causality.

Physical causality (Dowe 2000): A causal interaction is one that involves the exchange of a conserved quantity. For example, one billiard ball transmits momentum to another. Physical causality captures strikingly well what we often have in mind when we say that one thing caused another. This is presumably what Laplace (1814) had in mind when he wrote,

> We may regard the present state of the universe as the *effect* of its past and the *cause* of its future.

But physical causality is limited to direct physical interactions.

Interventionist causality (Woodward 2003 and Pearl 2000): X has a causal relationship to Y if and only if there is a possible intervention that changes the probability distribution of X, which results in a change to the probability distribution of Y. Interventionist causality is phenomenological. One can recognize an interventionist causal relationship even though one does not know the underlying mechanism.

Interventionist causality attempts to capture the intuition that if wiggling X results in Y wiggling, then X has a causal relationship to Y—or X can serve as something like a remote control for Y (Woodward 2014).

7.5.3.2 Switches

Switches stand a step removed from direct physical causality. Rather than transmitting a conserved quantity, a switch enables, disables, or redirects an energy flow. When one, say, flips a light switch, no conserved quantity is transferred from the switch or the switch flipper to the light. Yet flipping the switch satisfies the conditions for interventionist causation.

Switches illustrate the ubiquity of interventionist causality—from biology (e.g., gene switches) to computers (via transistors).

7.5.3.3 Symbolic Causes and Autonomous Causality

Symbolic causes are even further removed from direct physical causality. Consider these examples.

- Both (a) a court issuing an execution order and (b) the captain of a firing squad commanding *Shoot! cause* the death of the prisoner.
- Raising/lowering the price of an item *causes* the number of items sold to decrease/increase.
- A traffic light changing color *causes* cars to start/stop.

Although these causal relationships fit the interventionist paradigm, the causes are all symbolic. They produce an effect through the transmission of symbols.

What is a symbol? That's a very difficult question. I'll limit the discussion here to the following assertions.

1. Although represented physically, symbols are abstract. A bit value of 0 or 1 is the same symbol independently of how the bit is represented, e.g., by voltage level, magnetic moment, etc.
2. As an abstraction, a symbol cannot be the direct cause of a physical effect. According to one definition: an object is abstract (if and) only if it is causally inefficacious. –Rosen (2014)
3. A symbol's only property is that each symbol can be distinguished from all other symbols. In any symbolic system, the symbols can be interchanged without consequences.
4. Systems that manipulate symbols can associate meanings or values with them.

Given these limitations, how do symbols act as causes—especially of physical effects? To connect a symbolic cause to a physical effect requires a physical interpreter (a) with access to energy and (b) that can transform a symbol into physical action. We frequently call such interpreters agents. An agent's response to a symbol cannot be arbitrary; there must be a causal relationship. Yet different

agents may respond to the same symbol in different ways, and the same agent may respond to the same symbol in different ways at different times. Because an agent's response depends on the agent, I call this *autonomous causality*. Extraordinary! A symbol has a causal relationship to an agent; yet the agent determines what the effect will be.

Autonomous causality turns the tables on Laplacian causality. Laplace was talking about the laws of physics. Laplacian causality says nothing about the expected response to symbolic causes: *you don't expect a rock to respond to a traffic light*. Instead of causes pushing the world around according to the laws of physics, agents "choose"—by their internal organization—how to respond to symbolic causes.

An agent's design (or program) determines whether, and if so, how it will to respond to symbols. But designs can be changed:

- by designers when agents are explicitly redesigned;
- by natural or artificial evolution;
- by agents themselves that
  - change state as in a finite automaton.
  - change state to optimizes some measure, e.g., agents that "learn;"
  - run simulations before responding, i.e., agents that "think."

An agent's response to a symbolic input depends not only on the input but also on its history of past inputs. If one doesn't know an agent's possible states, its input history, or its mapping from input to state change, its response to a new input is unknowable and undiscoverable except by examining the output when that input is presented. It's almost as if each agent's interior is its own little universe.

But in some cases even knowing an agent's internal design and input history doesn't enable one to predict how it will respond. Consider AlphaGo, the computer program that beat the Go world champion. Its design and input history are known. Even so, the only feasible way to determine which move it will make is to give it an input and see what it does.

Switches and autonomous causation enable nature to create entities capable of producing complex and sophisticated energy flows—energy flows that may appear teleologically driven.

7.5.4 Autopoiesis and More–Mass Emergence

How do biological organisms hold together? The static structure of a biological organism is held together primarily[17] by negative interaction energy. But Schrödinger (1944) argued that

[17] Topological properties such as ball joints also help.

> present-day physics and chemistry cannot account for what happens within a living organism. All atoms are constantly in motion due to heat. Any lawfulness and orderliness that one might think of is made inoperative by the unceasing heat motion.

How much heat motion? Hoffmann (2012) compares it to a car in a 70,000 mph windstorm.[18] To defend against such disorder, biological organisms continually rebuild and repair themselves. In a review of autopoiesis[19] Luisi (2003) argued that all biological organisms have "a semipermeable chemical boundary within which they are capable of self-maintenance by self-generation of their components from within." *Autopoiesis* may be understood more generally to refer to all such self-maintenance activities, including the acquisition of resources and the avoidance of hazards.

Social systems—i.e., groupings of living organisms such as families, packs, societies, corporations, cities, countries, etc.—also hold themselves together through autopoiesis. When applied to social systems *autopoiesis* may be understood as self-maintenance within a self-created social boundary (Luisi 2014). The boundary need not be physical and may consist primarily of means to determine whether an entity belongs to an organization—e.g., a membership card or distinctive markings such as odor or color.

Because of their ongoing activities—which include the kinetic energy of motion—entities that rely on autopoietic mechanisms have *more mass* than the sum of the masses of their immediate constituents considered separately.[20] Autopoiesis allows nature to build entity types that would not persist otherwise.

7.5.5 *Biological Evolution*[21]

Evolution provides the fourth element in our framework for nature's open-endedness. Evolution increases the rate at which nature produces new entity types, and it helps nature create increasingly complex and sophisticated entity types.

[18] In his elegant book, Hoffmann points out that the "wind" is random rather than directional.

[19] The term *autopoiesis* has been dismissed as vacuous, trivial, overly complex, self-referential, circular, and intentionally mysterious. Maturana and Varela's original idea (1980) was to identify a category of systems that have the capacity to repair themselves. The claim is that all living systems are autopoietic but not that autopoiesis is sufficient for life. See Razeto-Barry (2012) for a review of the history and criticism.

[20] Abbott (2010) used the term *dynamic emergence* for systems that hold themselves together through self- maintenance. Abbott (2016a) argues that together static and dynamic emergence demystify emergence and render the term nearly superfluous.

[21] This section discusses biological-style evolution, not general evolution discussed earlier. For convenience, I'll use the term *evolution* for biological evolution and *general evolution* when referring to the broader process.

I will define evolution as a collection of processes involving entities:

(a) whose structure and function are characterized to a significant extent by meta-information that each carries with it, and
(b) which are capable of procreation,[22] an ability to produce offspring that resembles themselves. Procreation includes the transmission to the offspring of a possibly imperfect copy of the parent(s)'s meta-information.

In addition, evolution requires that entities be distinguishable and that whether an entity procreates depends in part on its individual properties. We say that an entity has a procreative advantage in a given environment if it has an increased probability of procreating in that environment. Given two entities each may have a procreative advantage in a different environment.

7.5.5.1 Evolution as a Driver of the Rate of Entity Type Creation

Evolution has a tautological flavor. Entities that succeed at procreating procreate. The more successful they are, the more offspring, similar to themselves, they produce. Evolution reflects an ongoing open-ended experiment in relative procreative advantage. Multiple entities occupy the stage simultaneously. Some procreate more successfully than others. But since procreative success depends on the environment, and since the environment may change with the differing relative success of various entities, what confers relative procreative advantage may change as advantages are achieved. Evolution becomes an ongoing search for what may be fleeting procreative advantages.

A procreative advantage often involves a new property. Consequently, evolution tends to increase the number of entity types. With new types come additional ways for other entity types to gain a procreative advantage, e.g., new types open the door to new predator or parasite types, leading to yet other new entity types. Evolution drives nature to create new entity types at ever increasing rates.

7.5.5.2 Evolution as a Means for Creating More Complex Entity Types

As defined, evolution requires that entities contain meta-information, which they pass on to their offspring. This feature has two important consequences: versioning and incremental improvement.

It is common commercial practice to produce new versions of products. Companies often produce product versions to fill various market niches. For different niches, a company might produce a less expensive version of a product, a more

[22] I'm using the term *procreation* rather than *reproduction* since the offspring are generally not exact copies.

luxurious version, a version with modified functionality, or all of these. Of course not every new version succeeds. But failure is accepted as the price of innovation.

How do companies create new versions? They tinker with the product design. The meta-information passed from parent to offspring serves as something like the biological equivalent of a product design.[23] In making often imperfect copies of meta-information, nature gets to tinker. Random tinkering with meta-information is far more likely to harm the recipient than to confer an advantage. But nature "fails fast." The bearer of a disadvantageous change dies. Every once in a while tinkering produces an incremental improvement, which survives and prospers. Via incremental improvement nature is able to create new versions of entities rather than being required to create new designs entirely from scratch.

As an illustration consider Zimmer's (2009) discussion of the evolution of eyes.

> Early in the evolution of animals, a serpentine protein, a class of protein in early eukaryotes that carried signals from one part of a cell to another, mutated to pick up light signals. Descendants produced light-sensitive eyespots packed with photoreceptors. These light-sensitive regions ballooned out to either side of the head, and later evolved an inward folding to form a cup. Early vertebrates could then get clues about where the light was coming from.
>
> A thin patch of tissue evolved on the surface of the eye. Light could pass through the patch, and crystallins, transparent proteins that can alter the path of incoming light so as to focus an image on the retina, were recruited into it, leading to the evolution of a lens. Natural selection favored mutations that improved the focusing power of the lens, leading to the evolution of a spherical eye that could produce a crisp image.
>
> The evolution of the eye did not stop there. Some species evolved the ability to see in the ultraviolet. Some evolved double lenses, which allowed them to see above and below the water's surface at the same time. Vertebrates adapted to seeing at night and in the harsh light of the desert. All those eyes were variations on the same basic theme established half a billion years ago.

Together negative interaction energy, the ability to create and tap stores of usable energy, autopoiesis, and evolution provide the basis for nature's constructive creativity.

Unfortunately, this picture's optimistic open-endedness may be limited. Autopoiesis and evolution require usable energy. Since the fate of the universe may be a heat death of maximum entropy the prospects for indefinite open-ended creativity look unpromising. But all may not be lost. Nothing guarantees that energy gradients will decrease faster than nature finds ways to exploit increasingly small differences.

[23] I wouldn't argue that DNA is the design of an organism or that DNA is the software an organism "runs." That's a greatly over-simplified picture of the role of DNA. But I think it's fair to say that the meta-information about the form and function of organisms that DNA encodes plays a role similar to a product's design.

7.6 Effective Field Theories

This section examines some examples of higher level phenomena that involve only interaction energy and considers how an effective field theory (EFT) approach might handle them.

Some physicists talk about EFTs as a way to explain higher level phenomena. An EFT considers interactions that occur within a particular scale and energy range. Other energy ranges are safely ignored since they don't contribute to the observed phenomena. Here's how (physicist) Nigel Goldenfeld (2011) put it at a conference of physicists and philosophers.

> Our ability to understand the physical world has to a large extent depended on the separation of scales that permits EFT descriptions to be useful. We can construct minimal models that enable efficient calculation of desired quantities, as long as they are insensitive to microscopic details. This works in many instances in physics. In other fields, such as biology, it is not so clear that these concepts are useful.

7.6.1 Example: Chemistry

In every chemical reaction the number of units of each atomic element is the same at the end as at the beginning. Why does this hold? The answer, of course, is that the forces that hold atomic nuclei together are much stronger than the forces in operation during chemical reactions, and they operate over a much shorter range. Consequently, atomic nuclei retain their identities during chemical interactions, and the forces responsible don't effect chemical reactions.

What makes the (chemical) conservation of elements a law (of chemistry) rather than just another consequence of physics? It's the coherence of chemical phenomena as an identifiable discipline. We find the world of chemistry to be both self-contained and distinct from other sorts of phenomena. It makes sense to talk about laws that hold within that world. Once one focuses on such a world, scientists develop a vocabulary for it. For example, one finds the following in Dickerson et al. (1979).

> The alkali metals are the most reactive known, and never occur naturally in the metallic state. Virtually any substance capable of being reduced will be reduced in the presence of an alkali metal.

Even though one explains the extreme reactivity of the alkali metals in terms of physics, alkali metals and redox reactions are not part of the vocabulary of fundamental physics.

Much of chemistry would seem to illustrate a successful application of an EFT approach.

7.6.2 Example: Geology

We understand the earth's geology in terms of a layered model that includes a crust, various layers of mantle, an outer and inner core, and transition zones between these layers. The crust consists, in part, of tectonic plates, which slide over deeper levels. As the plates move with respect to each other they store energy in the form of stress, which when released produces earthquakes, mountain ranges, etc.

Why does the earth consist of layers of material? Why do individual tectonic plates stick together and compress and then slip as they move past each other? Physics and chemistry can tell us. These phenomena are all consequences of interaction energy. Like chemistry, geology is a coherent science. But since the kinds of forces at play range broadly—from gravity to chemical electromagnetic forces and the electromagnetic forces that produce friction—it's unlikely that physicists will develop a geological EFT—except possibly for models at very high levels of granularity.

7.6.3 Example: Hydrodynamics

When submerged in a liquid (on earth) a substance will experience an upward pressure equal to the weight of the displaced liquid. Because of that, the weight, shape, and permeability of a displacing object will determine whether the object sinks or floats. A steel bar sinks in water. In the shape of a ship the same steel floats. In the same shape but perforated as a strainer, it sinks.

The physics of fluids explains the upward pressure. But ship building is a discipline of practical engineering. One must consider, among many other things, both (a) permeability and (b) the material properties that enable one to create water-impermeable bowl-like structures that hold their shape under various conditions.

Permeability depends on many factors including temperature and pressure. It is a function of electromagnetic forces and molecule sizes. It tends to be measured as a practical matter rather than computed as a theoretical result. The physical chemistry underlying the ability to form materials into stable shapes is a similarly practical matter.

We have been floating ships for millennia. It's unlikely we will see a shipbuilding EFT.

In summary, effective field theories exploit the commonsense notion that one should focus on what's relevant and ignore what isn't. They provide means for applying the mathematics of quantum field theory beyond the micro-world. They don't offer insights into nature as creative or constructive.

7.7 A Not Entirely Mathematical Regularity

In a discussion during (Carroll 2012), another conference of scientists and philosophers, Weinberg noted that summing the angles of any triangle on the earth's surface yields 180 degrees. He asked, rhetorically, why that was. In answering he pointed—to the surprise of many—to the relative weakness of the gravitational field at the earth's surface and hence the relative flatness of space.

Prior to 1916, general relativity would not have been part of the explanation. This geometric regularity would have been taken simply as a consequence of Euclid's axioms—along with an unstated assumption that space at the surface of earth satisfies those axioms.

This offers two lessons.

(a) Some regularities depend on mathematics and logic rather than on the laws of nature alone.
(b) When dealing with such mathematical regularities one must establish that the hypotheses upon which they are built apply in a particular physical situation.

7.8 Concluding Thoughts

Andersen and Hepburn (2016) credit medieval thinkers with defining *analysis* as: to examine a phenomenon to discover its basic explanatory principles. Anderson, Dawkins, Einstein, Schrödinger, Weinberg, and presumably most other scientists understand reductionism as synonymous with this sense of analysis. Scientists want to understand the underlying principles that explain how nature works.

But most philosophers of science have abandoned the attempt to characterize science-in-general in reductionist/analytic terms. Why? Most natural phenomena result from nature's constructive creativity. Although phenomena can be explained and understood through reverse engineering, a pure reductionist perspective tends to be blind to nature's constructive aspects.

The example of software helps us see why. The only reasonable way to understand software is constructively: how are software components put together so that the resulting compound produces the desired phenomena? Software developers have a significant advantage when creating new entities. The languages in which software is written includes as primitives mechanisms for putting components together. Software developers use these tools of construction in some sense for free.

Nature is not so lucky. To put physical things together requires physical mechanisms. Nature's constructive creativity depends on:

- negative interaction energy, the elementary forces that bind things together;
- means for accumulating and then spending stores of usable energy;
- autopoiesis, the strategy of applying usable energy to maintain the stability of structures that would not persist on their own; and

- evolution:
  (a) procreation and the transmission of meta-information (possibly imperfectly copied) and
  (b) competition for resources needed for autopoiesis and procreation.

References

Abbott, Russ. 2010. Abstract data types and constructive emergence. *Newsletter on Philosophy and Computers,* American Philosophical Society 9(2, Spring 2010): 48–56.

———. 2016a. The end of (traditional) emergence; introducing reactive emergence. *Journal of Public Policy and Complex Systems* 2/2: 91–107.

———. 2016b. *Autonomous causality*. Submitted for publication.

Andersen, Hanne, and Brian Hepburn. 2016. Scientific method. In *The Stanford encyclopedia of philosophy,* Summer 2016 Edition, ed. Edward N. Zalta. Stanford: Stanford University Press.

Anderson, Philip. 1972. More is different. *Science* 177/4047: 393–396.

Bricker, Phillip. 2014. Ontological commitment. In *The Stanford encyclopedia of philosophy* (Winter 2014 Edition), ed. Edward N. Zalta. Stanford: Stanford University Press.

Brigandt, Ingo and Alan Love. 2015. Reductionism in Biology. In *The Stanford Encyclopedia of Philosophy,* Fall 2015 ed, ed. Edward N. Zalta. Stanford: Stanford University Press.

Carroll, Sean. 2012. Discussions on emergence and reduction. In *Moving Naturalism Forward*. Sponsored by the Division of Physics, Mathematics, and Astronomy and the Moore Center for Theoretical Cosmology and Physics, California Institute of Technology.

Chomsky, Noam. 1966. *Topics in the theory of generative grammar*. The Hague: Mouton.

Dawkins, Richard. 1986. *The Blind Watchmaker*, 1986 ed. London: W. W. Norton & Company.

Dickerson, Richard E., Harry B. Gray, and Gilbert P. Haight. 1979. *Chemical principles*. 3rd ed. Menlo Park: The Benjamin/Cummings Publishing Company, Inc.

Dowe, Phil. 2000. *Physical causation*. New York: Cambridge University Press.

Dybjer, Peter and Erik Palmgren. 2016. Intuitionistic type theory. In *The Stanford encyclopedia of philosophy,* Spring 2016 edion, ed. Edward N. Zalta. Stanford: Stanford University Press.

Einstein, Albert. 1918. Principles of research. Address delivered before the Physical Society in Berlin at a celebration of Max Planck's sixtieth birthday. *Mein Weltbild*. Amsterdam: Querida Verlag, 1934.

Feynman, Richard. 1964. Knowing vs understanding. This is a segment of the complete series. Published on *YouTube*. See http://io9.gizmodo.com/watch-a-series-of-seven-brilliant-lectures-by-richard-f-5894600 for the complete series of lectures.

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

Fodor, Jerry. 1997. Special Sciences: Still Autonomous After All these Years. *Noûs*, Vol. 31, Supplement: *Philosophical Perspectives, 11, Mind, Causation, and World*, (1997), 149–163. Boston: Blackwell Publishing.

Goldenfeld, Nigel. 2011. Emergence and minimal models in condensed matter physics and biology. *Conference on Emergence and Effective Field Theories*, Perimeter Institute.

Hoffmann, Peter. 2012. *Life's Ratchet: How molecular machines extract order from chaos*. New York: Basic Books.

Koberlein, Brian. 2016. Death Spiral. *One Universe at a time* (blog).

Laplace, Pierre Simon. 1814 *A Philosophical Essay on Probabilities*. Trans. F.W. Truscott, and F.L. Emory. New York: Dover Publications (1951).

Loewer, Barry. 2009. Why is there anything except physics? *Synthese* 170/2: 217–233.

Luisi, P.L. 2003. Autopoiesis: A review and a reappraisal. *Naturwissenschaften* 90 (2): 49–59.

———. 2014. The minimal autopoietic unit. *Origins of Life and Evolution of Biospheres* 44 (4): 335–338.

MacLeod, M., and N.J. Nersessian. 2015. Modeling systems-level dynamics: understanding without mechanistic explanation in integrative systems biology. *Studies in History and Philosophy of Science Part C — Biological and Biomedical Science* 49: 1–11.

Maturana, Humberto, and Francisco Varela. 1980. Autopoiesis and cognition: The realization of the living. In *Boston studies in the philosophy of science 42,* ed. Robert S. Cohen, and Marx W. Wartofsky. Boston: D. Reidel.

McGuire, Brett A., P.B. Carroll, R.A. Loomis, I.A. Finneran, P.R. Jewell, A.J. Remijan, and G.A. Blake. 2016. Discovery of the interstellar chiral molecule propylene oxide (CH3CHCH2O). *Science*. doi:https://doi.org/10.1126/science.aae0328.

Nagel, Ernest. 1970. Issues in the logic of reductive explanations. In *Mind, science, and history*, ed. H.E. Kiefer and K.M. Munitz, 117–137. Albany: SUNY Press.

Pearl, Judah. 2000. *Causality*. New York: Cambridge University Press.

Razeto-Barry, Pablo. 2012. Autopoiesis 40 years later. A review and a reformulation. *Orig Life Evol Biosph.* doi:https://doi.org/10.1007/s11084-012-9297-y.

Rosen, Gideon. 2014. Abstract objects. In *The Stanford encyclopedia of philosophy,* Fall 2014 edition, ed. Edward N. Zalta. Stanford: Stanford University.

Schrödinger, Erwin. 1944. *What is life?* Cambridge University Press, reprint edition (2012).

Smart, J. 1959. Sensations and brain processes. *Philosophical Review* 68: 141–156.

Strassler, Matt. 2015. The energy that holds things together. In *Of particular significance: Conversations about science with theoretical physicist matt strassler*. Online blog.

van Riel, Raphael, and Robert Van Gulick. 2016. Scientific reduction. *The Stanford encyclopedia of philosophy,* Spring 2016 edition, ed. Edward N. Zalta. Stanford: Stanford University.

Weinberg, S. 2003. *Facing up*. Harvard University Press.

Wilson, E.O. 1998. *Consilience: The unity of knowledge*. New York: Knopf.

Woodward, James. 2003. *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

———. 2014. Scientific explanation. In *The Stanford encyclopedia of philosophy,* Winter 2014 edition, ed. Edward N. Zalta. Stanford: Stanford University.

Zimmer, Carl. 2009. The evolution of the eye. *The New York Academy of Sciences Magazine*, The New York Academy of Sciences. (October 9, 2009).

Chapter 8
Politics and Epistemology of Big Data: A Critical Assessment

Teresa Numerico

Verwisch Die Spuren
[. . .]
Was immer du sagst, sag es nicht zweimal
Findest du deinen Gedanken bei einem andern: verleugne ihn.
Wer seine Unterschrift nicht gegeben hat, wer kein Bild hinterließ
Wer nicht dabei war, wer nichts gesagt hat
Wie soll der zu fassen sein!
Verwisch die Spuren!
Sorge, wenn du zu sterben gedenkst
Daß kein Grabmal steht und verrät, wo du liegst
Mit einer deutlichen Schrift, die dich anzeigt
Und dem Jahr deines Todes, das dich überführt!
Noch einmal:
Verwisch die Spuren!
(Das wurde mir gelehrt.)

Erase Traces
[. . .]
Whatever you say, don't say it twice
If you find your ideas in anyone else, disown them
He who has signed nothing, who has left no picture behind
Who was not there at the time, who has said nothing
How are they to catch him!
Erase the traces!

Make sure, when you turn your thoughts to dying
That no gravestone divulges where you lie
With a clear inscription indicting you
And the year of your death, which convicts you!
Once again,
Erase the traces!
(That's what I was told.)
Berthold Brecht 1926

T. Numerico (✉)
Department of Philosophy, Communication and Performing arts,
University of Rome Tre, Rome - Italy
e-mail: teresa.numerico@uniroma3.it

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_8

Abstract In this paper I will discuss Big Data as a suite of new methods for social and political research. I will start by tracing a genealogy of the idea that machine can perform better than human beings in managing extremely huge quantity of data, and that the quantity of information could change the quality of the interrogation posed to those data.

In the second part of the paper I will analyse Big Data as a social and rhetorical construction of the politics of research, claiming in favour of a more detailed account of the consequences for its progressive institutionalization. Without a serious methodological assessment of the changes that these new methods produce in the scientific epistemology of social and political sciences, we risk to underestimate the distortive or uncontrollable effects of the massive use of computer techniques. The challenge is how to avoid situations in which it is very difficult to reproduce the designed experiment, and it is arduous to explain the theories that can justify the output of researches. As an exemplification of the problem I will discuss the work on emotional contagion led by Facebook and published on PNAS in 2014.

Until now it was difficult to explore all the Big Data projects' consequences on the perception of human intelligence and on the future of social research methods. The vision that there is no way to manage social data than to follow the results of a machine learning algorithm that works on inaccessible, epistemologically opaque and uncontrollable systems is rather problematic and deserve some extra consideration.

Keywords Big Data · Epistemology of social and political sciences · Machine learning · Epistemic opacity · Privacy · Control · Computational rationality · Complexity

8.1 Introduction, or a Proposed Genealogy for the Big Data

The idea to access the entire corpus of texts and content related to an object of study was conceived recently in connection with the development of the electronic digital technology, during the '60s of last century. Joseph Licklider published a book on the future of libraries (1965) in which he discussed the project of the direct interaction with the "fund of knowledge" in details. He is one of the pioneers of the network project, though he was not involved directly in the first practical steps of the creation of the Arpanet, the network that gave birth to Internet in 1969. In his book on *The future of libraries*, Licklider introduced the possibility of managing the entire "fund of knowledge" as a unique object of study that could be consulted in its integrity, even at distance. He hypothesized that the digital reorganization of libraries might be the vehicles that would facilitate a transformation in knowledge organization and, consequently, in knowledge acquisition.

Licklider suggested that the computer communication technologies would render possible a direct interaction between the 'fund of knowledge' and the result of the experiment designed by the researcher. He had in mind the classical scientific experiment, but the idea of Big Data was already there and could be applied to social and

political research as well as to physics and biology researches. It was based on the ingenuous representation of digitalization as a disintermediation. Licklider believed that the boundaries between the library where books and information were kept and the experimenter's laboratory forced the researcher to a cognitive mediation between the result of the experiment and the already acquired knowledge retained in books and other grey literature documents. He described the digitalization as a method that would allow a disintermediation between knowledge and the new experiment, avoiding the interposition of the scientist's cognitive structure, in order to interpret correctly the experiment.

Licklider's conviction was that the cognitive frame used by the scholar to make sense of the experiment's data was not needed if only the library could be merged with the laboratory. The transfer of the library within the laboratory was exactly his futuristic project – strongly pursued – from the moment (October 1962) he became head of the IPTO (Information Processing Technology Office), an office of the ARPA (Advanced Research Projects Agency, later called DARPA Defence Advanced Research Projects Agency) an agency of the US Department of Defence.

"In organizing knowledge, just as in acquiring knowledge, it would seem desirable to bring to bear upon the task the whole corpus, all at one time – or at any rate larger parts of it than fall within the bounds of any one man's understanding. This aim seems to call for direct interactions among the various parts of the body of knowledge" (Licklider 1965, 25).

According to Licklider, then, it was necessary to obtain 'direct interactions' between all parts of knowledge, and he was aware that, for a single human being, it was impossible to manage directly the necessary amount of information.

The machine should, then, act as a sort of expert colleague capable of giving the right advice to the scientist, and knowledge should be managed 'under human monitorship but not through human reading and key pressing' (Licklider 1965, 26).

The conclusion of his hypothesis seemed to be clearly formulated by Licklider himself:

> It no longer seems likely that we can organize or distil or exploit the corpus by passing large parts of it through human brains. It is both our hypothesis and our conviction that people can handle the major part of their interaction with the fund of knowledge better by controlling and monitoring the processing of information than by handling all the detail directly themselves (Licklider 1965, 28).

Another interesting consideration suggested by Licklider was that the human being could not be the unique and major agent in the process of acquiring knowledge, being more a sort of supervisor or coordinator of the effort of the machine's procedures. This proposal discussed a general issue as the future on knowledge, more than the narrower issue of the future of libraries. The machine was the only agent capable of interacting directly with what he called "the fund of knowledge". The explicit position of this thesis can be found in his words:

> He [the human being] will still read and think and, hopefully, have insights and make discoveries, but he will not have to do all the searching himself nor all the transforming, nor all the testing for matching or compatibility that is involved in creative use of knowledge (Licklider 1965, 32).

The more interesting element here is that he was not completely convinced that there would still have been a creative contribution of the human agent in the production of knowledge. He said: "he will [. . .] hopefully have insight and make discoveries", admitting that he was not sure that the human insight and the human ingenuity would still have maintained their relevance in the creative processing of information. If it were necessary to deal each time with the entire fund of knowledge in order to make new discoveries, no human being could be able to do it without the central role of a machine with the task of handling the data, using the adequate procedures and being programmed according to the most effective methods.

We can conclude then that the idea suggested here was connected with the Big Data approach, because it stated that it was better to deal with all the data potentially available, than to select only the relevant portion of those data. His hypothesis was that the activity of choosing and sorting out an original part of data, crucial for development of intuition and for the creative use of those data, was not achievable than using a machine. It was more effective to access all the data, subsuming it in a unique management procedure, in order to find the potential correlations among the data with the brute force of an exhausting research conducted by a machine. This hypothesis was not explicitly stated, but if we read between the lines it was clearly assumed as a starting point of the discussion, an a priori postulation.

Viewed in this perspective, the conclusion that Licklider suggested in 1965 seemed very similar to the objectives of the Big Data projects that started around the years '00 of this century. It could be considered as a genealogy of the ideological stance of the present technological design.

What Licklider could not foresee was the transformation that the new methods of data acquisition and organization produced on the practices of knowledge creation and on the methods of assessing results in all spaces where researches are conducted, not only in the laboratory, but also in the area of social, political and humanities studies.

8.2 Big Data and Their Rhetoric from a Critical Perspective

The idea surrounding Big Data is based on the same rhetoric point adopted by Licklider: storing, manipulating, interacting with the huge amount of data, now available in all fields of research ranging from physics, to biology, from social science to politics and humanities studies need to be handled directly by special machines, using algorithms specifically designed to treat the high quantity of data available. Such methods, though created by a specialized group of human beings, experts in data science, but not in the different fields from which data came, can only be monitored and controlled at a distance by human beings, because their brains do not exhibit the necessary plasticity and the needed amplitude to fulfil the task of dealing with all the available data, considering their volume, variety, and velocity.

This model that was called "3Vs" model (Laney 2001) was extended later to include the forth V "Veracity", that "refers to the level of reliability associated with certain types of data" (Schroeck et al. 2012).

The idea is that huge amount of data coming from different sources in continuous motion could be interpreted and managed by the new technologies invented for this specific new purpose in order to find new solutions for business problems and all sort of research issues.

The Big Data rhetoric prescribes that the methods used to make sense of the data could not be controlled directly by human beings because their volume, velocity and multifarious sources do not allow for a direct human intervention.

According to some scholars (see Barabási 2010; Mayer-Schönberger and Cukier 2013; Nielsen 2011) the access to the huge amount of data available in digital format will revolutionize the way scientific results are obtained also in the field of social sciences and humanities. This promise is very attractive to media companies that store all the data, but alarming for users, whose freedom is threatened, not only in terms of privacy. In this section, I will concentrate the discussion on the potentialities and risks of the use of Big Data solution for the social, political and humanities studies.

The collected digital traces left by almost any human activity in the almost entirely connected world, such as organizing a trip abroad, or starting a love affair, will allow researchers to manage not only statistical data on a population, but people's real lives. According to other scholars (Boyd 2010; Chun 2011; Gitelman 2013; Fiormonte et al. 2015), however, the excitement around the change of perspective of human sciences due to the manipulation of Big Data is completely overestimated. According to Boyd and Crawford "Big Data offers the humanistic disciplines a new way to claim the status of quantitative science and objective method. It makes many more social spaces quantifiable. In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth—particularly when considering messages from social media sites" (Boyd and Crawford 2012, 667). So it is imperative for scholars in the social, in the political and in the humanities fields to maintain their critical attitude towards those quantification techniques that seem to give their disciplines the appearance of objectivity. The reasons for the critical approach to Big Data are complex and various. One argument relates to the incompleteness and dirtiness of the data that form the basis of data-mining procedures. People are unaware that they are recording data on themselves when they participate in social networks, and, as a result, they may record false or incomplete information about themselves or their friends, which are then stored in the database and considered true. One of the reasons for this, according to Wendy Chun (2011, 93–94) is that people are always inconsistent in describing themselves, and any self-produced data design can only provide a misleading understanding of the subject and an inadequate prediction of his/her future preferences and actions. Others, like Jaron Lanier (2013), critique our current digital economy, making a case that links rising income inequality to the spread of what he calls "Siren Servers," or data-gathering companies:

> ... progress is never free of politics ... new technological syntheses that will solve the great challenges are less likely to come from garages than from collaboration by many people over giant computer networks. It is the politics and the economics of these networks that will determine how new capabilities trans- late into benefits for ordinary people. (Lanier 2013, 17)

Big Data raises a lot of critical issues relating to control and access to private information, as it is clearly shown by the data protection saga unleashed by the publication of documents by whistle-blower Edward Snowden and the New York Times, Guardian and other media during the summer of 2013. The details of the multi-million dollar programs managed by NSA (National Security Agency) and its British equivalent GCHQ (Government Communications Headquarters) show that PRISM and other tools are used with the (overt or hidden) help of the "big four" (Hotmail/Microsoft, Google, Yahoo! and Facebook). From email to texts, from mobile traffic to social network data, everything is collected and processed to prevent the potential risk of terroristic activities. The approach of English speaking intelligence agencies is based on the theory that it is better to know everything, than to miss information that may be relevant to a potential enemy action. The normal balance between the right to privacy and the right of executive power to protect society against violence has been completely subverted, given both the commonly misperceived level of risk in social networks, and the power of new brute-force decryption technologies to decipher formerly secret information. The opportunities offered by technology, together with the social perception of risks, has already changed the boundary between the permitted and illicit public exercise of power. According to O'Neil (2016) the quantitative attitude adopted in understanding many social behaviors is based on a manipulative and misleading displacement. The massive use of math and algorithms in understanding many different fields produces the consequent tendency to ignore all the aspects of social phenomena that are not passible of a quantitative representation, judging them as irrelevant, only because they are not measurable in an objective univocal way. O'Neal suggests that this tendency is causing a destructive effect on social justice and democracy.

8.3 Big Data and Power Issues: Facebook Data Science Team

Humanities and social studies offer a privileged perspective to assess the awareness that "raw data" does not exist (Gitelman 2013). Data creation is a necessarily biased activity as it is clearly stated in Bowker and Leigh Star (2000) relatively to the creation of categories and the organization of data into databases. There is no such a thing as a naked datum, without the possibility of perceiving it and arranging it into a structure of other data, that makes sense for a meaning creation project. Sometimes this organization is explicitly performed and implied in a theoretic frame, which is consciously explained by the scientist who is performing the experiment or is analyzing its output. Other times, as it often happens in Big Data experiments, the 'theory' is buried under a bigger stratum of data, so that it is a bit more difficult to extrapolate it, in order to evaluate its plausibility and its heuristic strength.

According to an often-cited article by the celebrated Chris Anderson on Wired (June 4, 2008), dedicated to the "end of theory", the massive availability of data through search engines makes the scientific method of research, with its hypotheses, theories and experiments, effectively obsolete. Computers might be better placed to explain the vast amount of data collected and stored in various databases. The paper was written before the huge success of Social Networks and is concentrated only on Big Data activities performed by search engines, but it is very easy to update the position by including all the Data science experiments in social, political and humanities studies that obtained so much public recognition during the last decade. Anderson cited Peter Norvig, Google's director of research, who said, after George Box, "All models are wrong, and increasingly you can succeed without them." According to this viewpoint, "correlation is enough. We can stop looking for models. We can analyze the data without hypotheses about what it might show" (Anderson 2008). This techno-fundamentalist position can only lead to dismay, but we need some little child available to announce that the emperor is naked and ridiculous in front of his subjects, otherwise we will not have any exit strategy to our scientific future based on data science experiments performed directly by the sophisticated machines, and their even more refined algorithms, whose characteristics are unknown even to the scientists who should manage the experiments. Today more and more funds are redirected towards planning technological infrastructures, while investments in research laboratories and the like are being reduced.

In June 2014 PNAS journal published the description of a Facebook experiment on measuring emotional negative and positive contagion by altering the news feed of 689,003 English users.

The paper (Kramer et al. 2014) was written by Adam Kramer (at that time core data science team Facebook) and signed also by two scholars in social sciences who worked at the Dept. of Communication and information science, Cornell University (see Schroeder 2014 for a complete analysis of Facebook experiment). The conclusion of the experiment according to the authors was:

> We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues (Kramer et al. 2014, 8788).

The project of the experiment was to test the possible emotional contagion via the newsfeed, so Facebook altered the newsfeed for users in order to validate the thesis according to which emotional contagion can happen with distance contacts and – researchers affirmed – it is valid both for negative and positive emotions.

In order to define positive or negative posts they used words as negative and positive according to a linguistic tool: the Linguistic Inquiry and Word Count software (LIWC 2007). This software defines words as positive or negative without taking into account the context in which they are used. They considered positive posts with at least a positive word and negative a post that included at least a negative

word. The researchers did not analyze the real posts but only the definition obtained according to the software. The first discussion that we need to open is: how is it possible to define a text as positive or as negative in terms of emotional expression, without any contextual evaluation? Is it safe to quantify the messages counting positive and negative words that are used? What if the posts were sarcastically conceived? And what if they were just a joke? How is it possible to quantify the emotional impact of a joke?

According to the researchers they assess two different results: one was relative to the lack of necessity of non-verbal behavior for the emotional contagion to take place. "Textual content alone appears to be a sufficient channel" (Kramer et al. 2014, 8790).

And secondly that: "emotions can spread throughout a network, the effect sizes from the manipulation are small" (Kramer et al. 2014, 8790).

Even the authors were conscious of the minimal size of the change (d. = 0.001), but they were confident that the mimicry effect would be relevant also for 'public health'.

My belief is that the rhetoric around the findings has to be taken into account, also considering that a prestigious journal such as PNAS accepted to publish the results even if there were various and different biases in it.

First of all we need to consider the issue of informed consent. How can you ask the informed consent for such a type of experiment to the users? In fact there was no informed consent given by the users, while Facebook was experimenting on their friends' newsfeed. After facing so many critics about the design and conduction of the experiment, Facebook changed the privacy policy by saying that between the terms and conditions that users accept when they subscribe, there was also the acceptance of experiments with the newsfeed.

The question is not only related to the specific case of Facebook experiment on emotional contagion but more generally on all the experiments that involve the use of personal data for obtaining conclusions related to habits, preferences, or attitudes of people, by analysing their personal public information. According to (Custers 2016) it would be useful to have an expiry dates for informed consent, so that it is would be not allowed to use personal data after a certain time from the explicit informed consent to access and manipulate personal data.

There was also a discussion about the participation to the experiment by the two Cornell University researchers, but the IRB (Institutional Review Board) of the university accepted the fact that the responsibility was of Facebook. The authors stated:

> [The work] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that: "Because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Re- view Board] determined that the project did not fall under Cornell's Human Research Protection Program. This statement has since been confirmed by Cornell University (PNAS, 10779)

This is relative to the ethical and legal aspect of the experiment. Who is in control of the data during the experiment? Who is responsible for the data and the conduct during the experiment? Do we believe that Facebook is committed to correctly handle data, without ethical, legal and epistemological problems?

Users tested in the experiment did not receive any prior information or opt-out opportunity, because Facebook is a company and not a research institution, there was no need to ask for any extra consent than that which was already granted by signing the service agreement.

The defence of Facebook with respect to this point is based on the fact that the company always manipulates user experience (Boyd 2014). The manipulation of the newsfeed is not a special characteristic of the experimental context. Facebook team uses to manipulate the newsfeed for all sorts of reasons, the amusements of users, the increase of their permanence within Facebook, as well as the advertisement system, the experiments on which are the most effective posts. So manipulation was not a special case of this specific experiment.

The assumption that laid behind this perspective is that once data is anonymous, it is not related to social researches on individuals. So it is not necessary to have the special attention that is attributed to social research, relative to individuals. There was a lot of discussion on this point within the field of biomedical research, as it is suggested in (Metcalf and Crawford 2016). However it is very difficult to affirm that data are completely anonymized as it is shown by various famous investigative results on anonymized data. Metcalf and Crawford cite the case history of the investigation on the real identity of the street artist Bansky. Moreover it is not true that anonymous personal data crossed with other personal data related to the same people are not classifiable as social research related to individuals.

According to Metcalf and Crawford (2016) "any move to exclude data science research from review, and more broadly, to consider it outside of human-subjects research, is thus premature and potentially dangerous" (p. 10).

Ethical questions and the discussions for a correct treatment of data within the area of Big Data related to social researches are not the only critical issue to take into account.

The recently created research field of critical data studies should deal with all these issues trying and obtaining consensus on the fact that it is crucial to have a "deeper understanding of data subjectivity, including an account of the fundamental responsibility that researchers have to care for the well-being of their subjects" (Metcalf and Crawford 2016, 10).

Data subjectivity in the case of Big Data related to personal data raises not only ethical and legal questions but also political questions that need to be addressed in order to find a fair balance between the rights of researchers and of citizens and digital service users to be protected, correctly informed and not expropriated of their personal information for business reasons.

Moreover there is a question related to the propriety of the data and to the possible controllability of the experimental results. Who owns 'scientific' data? Facebook is a private company and decides internally about control practices, but there is no sign that the internal data science team want to follow the regular rules

followed by research institutions. My concern is also that the scientific journals tend to be less restrictive with their peer review policy if the authors come from an important Internet company.

As suggested by Ahonen we have to be careful because:

> should this institutionalization [of Big Data methods] take place, it would also generate its 'rationalized myths' (Meyer and Rowan 1977) with the exaggeration of the merits and contributions of these methods and formal rather than substantive commitment to them. We would also witness not only the enhancement of the rationality of the core analytic processes of research by means of the Big Data methods, but also the strengthening of the external legitimation of institutions of research by the same means (Ahonen 2015, 8)

Following the suggestion of Zwitter (2014), it is necessary to acknowledge that there is a shift in traditional ethical attitude towards responsibility and agency in the era of Big Data because a non human agent and a collective non human agent cannot be considered responsible because they do not possess the completeness of moral agency in terms of knowledge, freedom of choice and causal connection between an action and its consequences. In this context we fail to consider Big Data algorithms responsible for the previsions and for the connections they make explicit, though they create a clear unbalance of power between Big Data collectors and Big Data generators. Analyzing the relations between the different stakeholders (collectors, utilizers generators of Big Data), Zwitter (2014) pinpoints that there is a very different power relations between them in terms of individual agency. The generators of the data cannot control the use of those data and also they cannot give any informed consent for the use of the data. According to Zwitter this situation of disparity among agents "changes foundational assumptions about ethical responsibility by changing what power is and the extent we can talk of free will by reducing knowable outcomes of actions, while increasing unintended consequences" (2014, 3).

8.4 Epistemological Opacity and the Power of the Algorithms

According to Alan Turing, one of the fathers of the idea of the machine intelligence, if machines could exhibit intelligent behaviours, they should, at the same time, commit some mistakes from time to time, as well as human beings. The ratio for this thesis is that in order for behaviour to be intelligent, it is necessary that it is creative, but creativity and originality are risky, and potentially exposed to failure. "There are indications however that it is possible to make the machine display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess" (Turing 1945, p.16). This thought was often repeated in his works on machine intelligence. See for example:

> I would say that fair play must be given to the machine. Instead of it sometimes giving no answer we could arrange that it gives occasional wrong answers. [...] If a machine is expected to be infallible, it cannot also be intelligent. There are several mathematical theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility (Turing 1947/2004, 394).

Turing's approach towards machine intelligence was based on the assumption that the machine had to mimic human intelligence, and if it were successful then, as in the case of the human beings, intelligent performances were not always exceptional. They were, to say the least, variable. Human beings are used to make a lot of mistakes, but from time to time they could be very creative and solve a very difficult problem showing a great deal of ingenuity and wit. Turing thought that an intelligent device should obtain similar results: in order to be sometimes surprising, the machine should be allowed to make mistakes.

If we look at the present scenario of big data results, however, there is a very different perspective in action when we want to attribute intelligent results to machine learning algorithms that make sense of big quantity of quantified data.

Following this different scenario, some of the critics about big data epistemology adopt a different view of artificial devices knowledge practices. The scholars that follow this path apply to those new research methods the traditional discussion of philosophy of science, with special regards to epistemic opacity. Some authors are convinced that:

> Instead of focusing exclusively on the potential consequences of the Big Data phenomenon, we can gain additional insight from examining its social and political, but also its technical and epistemic roots (Rieder and Simon 2016).

I completely agree with this vision, whose aim is the objective of this section. According to (Symons and Alvarado 2016) in particular, the quantification of errors and of error propagations in Big Data methods should be taken into account properly in order to avoid dangerous mistakes that are difficult to detect, due to the structure of machine learning algorithms used within these data management tools. Following Humprheys (2004, 2009), they affirm that the massive use of computer within Big Data methods poses new epistemic questions due to the fact that the computational rationality cannot be compared with human rationality. Humprheys (2009) affirms that "Here a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process" (pp. 5–6).

Now it is possible to confute this argument by saying that it is not true that human beings are not in control of the algorithms of machine learning used in Big Data analysis. However if this is the reply, we have to face a contradictory vision. We need to find an agreement between the expectations of unforeseeable results as output of machine learning algorithms analysis of Big Data archives and the fact that the programmers are in control of the algorithms they use. If they are in control, then they know what they are looking for and how to get there.

In this scenario (which it is likely to be the case) then it is impossible to sustain the thesis that there is no theory behind the choice of Big Data methods. The theory is opaque and invisible, because it is embedded in the software chosen to program the machine learning tools that are used to make sense of Big Data under consideration. But the theory is there. It is an invisible layer that is implied by the algorithms decision strategies, chosen, maybe even unconsciously, by the programmers.

We are facing a dilemma here. In the hybrid epistemic systems, which are used to manipulate huge amount of data, human cognitive abilities are not enough to obtain useful results. We are substituting human capabilities with machines' characteristics that – unexpectedly, considering Turing's thesis – did not mimic human intelligence. Mechanical devices follow, instead, different procedures, and above all, favour different strategies to organize and categorize Big Data, received as input. Machines are, in fact, capable of manipulating a huge amount of data, impossible to manage for human cognitive instruments.

However they are not neutral with respect to bias and prejudices that they apply to the research domain. Machine learning techniques follow an experimental question that they want to test, finding some types of correlations between data, so we can affirm that the research is not neutrally shaped.

According to Burrell (2016), it is important to raise a flag and evaluate carefully the level of epistemic opacity "that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully". The paper tackle the issue with a very technical approach, it analyses computer science and industries practices and it gives the results of some coding manipulation experiments in order to argue in favor of the thesis that there are some theoretical elements inside the choices of the software methods, and those choices are biased according to the expected experimental results. According to Burrell 2016 software is produced following research models, no matter which is the preferred technique adopted. This conclusion is particularly relevant when applied to the continuous increase in the scale of experiments. "The continuing expansion of computational power has produced certain optimization strategies that exaggerate this particular problem of *opacity as the complexity of scale* even further" (Burrell 2016, 9, emphasis in the paper).

We share Burrell's concerns about "lack of 'fairness' and discriminatory effects [. . .] of the algorithm's objectivity"(p.9). Her uneasiness with the opacity of the algorithms is not only relative to the level of secrecy of software platforms, which deliberately keep their methods obscure, as suggested by (Pasquale 2015). The conclusion of Burrell's work suggests that the opacity would still remain there, even if all the corporate secrecy were dissolved by regulatory impositions (Burrell 2016, p.10).

Technical opacity seems to be connected with the rising risks of biases and inequality in processes controlled by Big Data machine learning algorithms. The arguments in favor of this perspective are the central issues of the research conducted by Cathy O'Neil (2016). In her book, *Weapons of math destruction*, O'Neil discusses about a very crucial point: the ignored presence of theory and prejudices embedded in algorithms and other math tools adopted to deal with big quantity of data. She was one of the financial quants that after the 2008 huge crisis developed the feeling of being responsible for all the terrible economic disasters that affected weak people so badly, due to the collapse of part of financial system. She quitted and, after a reflection period, she decided to dedicate her effort to the development of the public awareness of the unfairness of many of the mathematic tools chosen to create correlations between data. According to her "data scientists

all too often lose sight of the folks on the receiving end of the transaction. They certainly understand that a data-crunching program is bound to misinterpret people a certain percentage of the time [. . .]. Their feedback is money which is also their incentive. Their systems are engineered to gobble up more data and fine-tune their analytics so that more money will pour in" (O'Neil 2016, pp. 12–13).

In her opinion, epistemic opacity has not only to do with secrecy, or with the neutrality presuppositions about technology but also with the fact that the objective of Big Data systems and algorithms is not to understand reality of society or whatever else they apply to, but to make more money. Money is the central reason for their development and the unique scope of their activities. The theory behind the tools is present, even if it is obfuscated and maybe somehow unconscious, but the methods adopted push towards specific results, whose results benefit some of the actors and damage others.

8.5 Mistakes and Reproducibility

The epistemic, economic and political opacity described in the previous section leads to the open issue of error detection. Is it possible to detect errors and misclassifications when the mechanism behind the categorization is opaque and difficult to understand, because it follows unwritten, and maybe also implicit rules?

There are some cases, however, in which the Big Data procedures are not the unique devices to deal with the same data. Google Flu Trends (GFT) is a relevant example to clarify the error detection problem with Big Data. In 2013 GFT failed to foresee the peak of flu season by 140 percent (Lazer 2015). In 2014 there was a paper published on *Science* that explained why (Lazer et al. 2014). The investigation of the research group was possible because the flu peak was measured also using small data statistical techniques by the Centers for Disease Control and Prevention (CDC). Those data allowed a prevision with a shorter time lapse, but still allowed to count the emergencies and offer an alternative quantified model of the phenomenon of the flu peak. So it was possible to replicate the experiment, using different methods discovering the deviation of the anticipated prevision. In this case, the evaluation procedure for the output of the algorithm was controllable, by adopting a different technique, which is not always the case.

In the *Science* paper, researchers made some suggestions about why the correlations failed so deeply. "they [Google team] were remarkably opaque in terms of methods and data – making it dangerous to rely on Google Flu Trends for any decision-making" (Lazer 2014). They used a lot of seasonal terms that were included in the list of meaningful terms. This is the result of the unproven prejudice that in deep winter epidemic of flu becomes more likely to happen. Moreover following the idea of the self-fulfilling expectation after introducing the method and choosing the target terms they render those terms more easy to use in the term suggestion list, increasing the number of research "throwing off GFT's tracking" (Lazer 2014) .

Another problem was the feedback effect. Once that the method is put into place, which are the control flags and feedback effects that monitor the efficiency of the prevision strategy? When Google evaluates the effectiveness of the formulation of a text ad in Adwords, the feedback is immediate because it monitors the behaviours of millions of people exposed to the ads. But when the output previsions are an offline effects, Google has fewer ways to monitor what happened in reality. It chooses to keep on using the same established and trusted methods, without understanding the invisible vicious circle that risks to bias the results of the algorithm adopted.

Algorithmic opacity and the rhetoric of the uselessness of theory are the principal responsible for the lack of controllable procedures and of the effectiveness of experimental tools adopted.

The problem raised here is relative to the general problem of reproducibility and controllability of results and experiments. In an editorial published in May 2016 on *Nature* Reality check on reproducibility (2016), it is suggested that there is a potential crisis of reproducibility in science in general. In *Nature* survey two third of the researchers and readers responded that "current levels of reproducibility are a major problem". This is true all over scientific practices and methods and in particular in sciences that involve the evaluation of behaviours of human beings (see for example Nosek et al. 2015). However this is particularly true when the techniques are opaque and the data proprietary as in many situations in which Big Data tools are used.

So we need to improve reliability and methodology of Big Data setting standards and assessing methods as suggested by Raghavan (2014):

> Machine-learning methods are a valuable part of our toolkit in understanding behavior, but we do not yet understand the precise limits of their applicability [. . .]
>
> The biggest contributions before us are not new algorithms or new social theories but new methodologies for decomposing hard questions in the social sciences into a series of robust analyses that are replicable and composable (Raghavan 2014).

As suggested by Gillespie (2014) algorithms that are use with Big Data come together with a database that needs to be organized by them: "before results can be algorithmically provided, information must be collected, readied for the algorithm, and sometimes excluded or demoted" (2014, 169). So this means that we cannot understand the algorithm in isolation with respect to the data that it is supposed to make sense of. This is another layer of complication in order to understand the functioning of the complex system that is determined by the interaction of the data arranged so that the algorithm can work on it.

The level of possible mistakes, manipulations and misunderstanding increases the layers of possible mystifications and incomprehension of the phenomenon represented by the data.

Interpretation instance is hidden between these strata of multiple representation and organization of data, without the awareness of the researchers. But it still there performing the biases and the prejudices embedded in the cleaning and organizing of data as well as in the anonymous and deceptively neutral rules of the procedure implemented by the algorithm.

One of the key elements that is worth our attention about big data and their organization algorithm is that we need to understand: "For whom, besides insurance companies, is this correlation – the revelation regarding mutual habit formation – useful? These studies [. . .] are not designed to foster justice" (Chun 2016, 14–15).

8.6 The Risks of a General Archive of All Possible Data

The organization of all human knowledge – the task of Big Science enterprises – as well as being extremely difficult, has also potentially alarming side-effects. It is clear that the ambition of the search engines and social networks is to become the archive of the Web or even better the archive of the walled gardens in which the Web is about to be transformed, or has already been transformed, without our consciousness or our consent.

Instead of the multifarious and multifaceted area of the world wide hypertext, conceived at the beginning of the WWW invention, we risk to open our eyes in a place full of burdens and biases, in which corporations own all the data and the produced multimedia content. The cyberspace, viewed from this perspective, can be compared to the worst dystopian novels that we could conceive. The user is not only relevant for the system because of its attention capacity but also because his/her habits, desires, preferences feed the system of information, of which it is hungry. Are we only the public of the online media or directly the product that is sold to advertisers and to the other users in a cannibalized environment? The users seem to be the subject of the control strategy and the object under control. Our attention and the consequent data, which can be collected from our active participation to the online universal chat is the high stake of these dangerous games, whose rules are opaque and unknown: who is the arbiter and who is the irrelevant puppet?

We need to assess that the procedure of building a multilayer archive of intentions, habits, desires, preferences, beliefs, convictions is not neutral activity in which information can be collected without order or rules. Whoever constructs such an archive interprets that information and establishes both what can be found and retrieved and what will be irredeemably hidden and buried, regardless of the genuine intentions of the archivist or the reader/retriever. In short, the archive constructs the meaning of phrases that would otherwise have no organic structure. As Foucault writes in his Archaeology of Knowledge:

> The archive is also that which determines that all these things said do not accumulate endlessly in an amorphous mass, nor are they inscribed in an unbroken linearity, nor do they disappear at the mercy of chance external accidents; but they are grouped together in distinct figures, composed together in accordance with multiple relations, maintained or blurred in accordance with specific regularities. (Foucault 1969/1982, 145–146)

Therefore, the archive is the horizon of meaning that determines the possible knowledge of events, ideas or people. It fixes the regularities that allow us to interpret, in each moment, the world around us and to establish what information survives and what will disperse as mere noise, by losing access to a defined and

organized form. Far from being a dusty and forgotten place, the archive in all its forms is the beating heart of a civilization. The work of the search engines must be connected to this sphere, and it is clear that humanists should supervise the criteria, principles and "regularness" adopted by these technological instruments. Search engines, big retailers, social networks expressly declare that they want to take on the role of being the super-archives of all online knowledge, and ultimately, of all knowledge, period.

According to Foucault, "The archive cannot be described in its totality" (1969/1989 147). It is clear that it needs an external world, an outside to refer to: there cannot be an archive without an outside-the-archive (Derrida 1995). It is just this outside-the-archive that we risk losing, unless we retain the critical spirit and vigilance of the humanities, and prevent technologies from taking over the spirit of research, by permitting a mechanical rule like the ranking algorithm, no matter how efficient, to pass unquestioned.

An interesting study shows the manipulative capacity of search engine on undecided voters during elections in countries where the use of the network is massive (Epstein and Robertson 2015). The outcome of the experiments was very frightening and supported the evidence of a great uncontrolled influence of search engines on elections, politics decisions and public opinion biases.

According to the preoccupied authors:

> Given that search engine companies are currently unregulated, our results could be viewed as a cause for concern, suggesting that such companies could affect—and perhaps are already affecting—the outcomes of close elections worldwide. Restricting search ranking manipulations to voters who have been identified as undecided while also donating money to favored candidates would be an especially subtle, effective, and efficient way of wielding influence.
>
> Although voters are subjected to a wide variety of influences during political campaigns, we believe that the manipulation of search rankings might exert a disproportionately large influence over voters.

If we accept their conclusions, it is necessary to underline a clear alarming situation that has to do with the power of unregulated gatekeepers on common citizens. Internet users deserve a more respectful treatment that it is likely that this result is reachable only with an international well-organized regulatory activity, as suggested also in (Pasquale 2015).

8.7 Epilogue

The phenomenon of the quantified self will have the consequence to consider irrelevant all those human characteristics that cannot be measured and represented as "data". This habit of measuring any single beat of our heart, as well as the rate of nail growing, or of our daily jogging performing, the substances needed from our skin to stay young forever, etc. is transforming our life in a constant succession of processes of measurement, with a special attention to miniaturizing

and optimizing the feedback tools. Google Glasses was a valid, through abortive, example showing which is the trend in action. We are more and more familiar with this quantification attitude so that we acquiesce to accept that all human characteristics can be measured easily. However we are not conscious that this attitude towards quantification of all personal phenomena produces an involuntary, menacing output: once assessed and reduced to 'objective' data the quantification process does not need any careful management, while it can be used to create a facile scale on which we can project all the people in a long unequivocal line that connect the first with the last of the queue.

Understanding of any phenomenon is thus increasingly connected with the production of the data that pretend to understand our environment and ourselves. But as suggested by Geoffrey Bowker "getting more data on the problem is not necessarily going to help" (Bowker 2013, 171). While it would be better to admit that "embracing the complexity of inquiry as a generative process of collaborative remix can push us to accept that no matter how good our tools, algorithms, or filters, we cannot possibly explain the whole of any situation" (Markham 2013, 10). We need to refuse being blackmailed by the objectivity and complete measurability of phenomena within the data-program-data cycles (Bowker 2013, 170) and start exercising our "strongly humanistic approach to analyzing the forms that data take; a hermeneutic approach which enables us to envision new possible futures" (Bowker 2013, 171).

There is no easy solution to the problems raised by the information society and its tools of knowledge control, such as search engines, social networks, and all the other social apps that produce Big Data mechanisms and systems. The only possible antidote is to increase the standards and the intensity of education in critical thinking and e-literacy; in order to encourage the development of a multiplicity of sources and the skills needed to enquiry social phenomena.

The risk of a politics based on Big Data is that the biopolitics of power will be based on data that nobody knows in details, all health decisions, traffic decisions, legislative behaviours could be taken on the basis of a big black box, whose procedure is opaque and whose logic is unknown by citizen, while their lives are managed by the Big Data system (Pasquale 2015). We cannot invoke privacy protection to defend us, because collected data are not tracing us by our single individual electronic footprint, but only as a member of a group that shows a specific distinguished behaviour. According to Byung-Chul Han (2013, 87–102) the transparency society is similar to a surveillance society, because instead of trust there is control. Instead of Big Brother there is Big Data, where our entire life is protocolled. Citizens of digital panopticon feel that they are free but it is just an illusion. According to Han we are in the psycopolitics age which has now overcome the biopolitics era. At present, psycopower and distant control are able to program us. Data mining, in Han's opinion, would control people because the system is in touch with the collective digital unconscious, being capable of influencing people directly by engraving its influence within the inner feelings and emotions of individual users.

Han's approach, though very refined and deep, is founded on the attribution of a disproportionate power to digital devices, like many techno-antagonists. He is partly influenced by Martin Heidegger vision of technique. We can also assume the point of view of Wendy Chun (2011), according to which the database is always inevitably full of mistakes and noise, so there is no system that is capable of representing our habits, desires and buying behaviours with a reasonable possibility of genuinely resembling us. The authentic challenge is to avoid the false attribution of infinite power to the machine; the only risk is that the illusory vision often overcomes the materiality of experience.

The anomaly that should alert us is that, while everything is supposed to be transparent, companies that deal with data, like Acxiom or Google itself, for example, work in a very protected environment and keep a very strict secret about procedures adopted, hiding even the locations of the laboratories where data is processed. Also Rodotà suggested that the strict protection of the secret procedures within Google laboratories characterizes its power signature (2014, 38). We have to pay attention to the transformation of authority that is the output of the blind trust that we are ready to confer to communication technology management.

We have to understand the trends of the present orientation toward knowledge production using Big Data methodologies and decide what we are ready to accept and what we want to discuss. We have to build a secular attitude when confronted with technology-oriented decisions. Machine is not our next religion and we should keep a critical positioning when faced with the outputs of techno-power.

References

Ahonen, P. 2015. Institutionalizing Big Data methods in social and political research. *Big Data & Society* 2(2, July). http://bds.sagepub.com/content/2/2/2053951715591224. Accessed 9 Oct 2016.

Anderson, C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 23 June 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory. Accessed 9 Oct 2016.

Barabási, A.L. 2010. *Bursts: The hidden patterns behind everything we do, from your e-mail to bloody crusades*. New York: Dutton.

Boyd, D. 2014. It's complicated. In *New Haven*. London: Yale University Press.

Bowker, G.C. 2013. Data flakes: An afterword to 'Raw Data' is an oxymoron. In: Gitelman. 2013, 167–171.

Bowker, G.C., and S. Leigh Star. 2000. *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Boyd, D. 2010. Privacy and publicity in the context of Big Data. *Raleigh* (nc), 29 April 2010 http://www.danah.org/papers/talks/2010/WWW2010.html. Accessed 9 Oct 2016.

Boyd, D., and K. Crawford. 2012. Critical questions for Big Data. *Information, Communication & Society* 15 (5): 662–679.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1, January). http://bds.sagepub.com/content/3/1/2053951715622512. Accessed 8 Oct 2016.

Chun, H.K.W. 2011. *Programmed visions: Software and memory*. Cambridge, MA: MIT Press.

Chun, W.H. 2016. *Updating to remain the same*. Cambridge, MA: The MIT Press.

Custers, B. 2016. Click here to consent forever: Expiry dates for informed consent. *Big Data & Society* 3(January–June). http://bds.sagepub.com/content/3/1/2053951715624935. Accessed 9 Oct 2016.

Derrida, J. 1995. Mal d'archive: une impression freudienne. Édition GalilÉe, Paris; En. Trans. Archive fever: A Freudian impression. Chicago: Chicago University Press, 1998.

Epstein, R., and R. E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In Proceedings of the National Academy of Sciences of the United States of America, PNAS *112*(33): E4512–E4521. https://doi.org/10.1073/pnas.1419828112. Accessed 9 Oct 2016.

Fiormonte, D., T. Numerico, and F. Tomasi. 2015. *Digital humanist. A critical enquiry*. New York: Punctum Books.

Foucault, M. 1969/1982. *The archaeology of knowledge*. London: Vintage.

Gillespie, T. 2014. The relevance of algorithms. In *Media technologies :Essays on communication, materiality, and society*, ed. T. Gillespie, P. Boczkowski, and K. Foot. Cambridge, MA: The MIT Press.

Gitelman, L., ed. 2013. *"Raw Data" is an oxymoron*. Cambridge, MA: MIT Press.

Han, B.C. 2013. *Im Schwarm : Ansichten des Digitalen*. Berlin: Matthes & Seitz.

Humphreys, P. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. New York: Oxford University Press.

———. 2009. The philosophical novelty of computer simulation methods. *Synthese* 169 (3): 615–626.

Kramer, A.I. et al. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(24, 17 June): 8788–8790. www.pnas.org/cgi/doi/10.1073/pnas.1320040111. Accessed 9 Oct 2016.

Laney, D. 2001. *3D data management: Controlling data volume, velocity, and variety, meta group (now Gartner Group) report*. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed 9 Oct 2016.

Lanier, J. 2013. *Who owns the future?* New York: Simon & Schuster.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. 2014. The parable of Google Flu: Traps in Big data analysis. *Science* 343(14 March): 1203–1205.

Lazer D., Kennedy R. 2015. *What We Can Learn From the Epic Failure of Google Flu Trends*. Wired, 10/1/2015. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. Accessed 2 nov 2018.

Licklider, J.C.R. 1965. *Libraries of the future*. Cambridge, MA: The MIT Press.

LIWC (2007) Linguistic Inquiry and Word Count (LIWC) program dictionary, 2007 version http://liwc.wpengine.com/. Accessed 2 Nov 2018

Markham, A.N. 2013. Undermining 'data': A critical examination of a core term in scientific inquiry. First Monday, 18, 10, October. http://firstmonday.org/ojs/index.php/fm/article/view/4868/3749. Accessed 9 Oct 2016.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A revolution that will transform how we live, work, and think*. London: John Murray.

Metcalf, J., and K. Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data and Society* 3(January–June). http://bds.sagepub.com/content/3/1/2053951716650211. Accessed 9 Oct 2016.

Meyer, J.W., and B. Rowan. 1977. Institutional organizations: Formal structure as myth and ceremony. *American Journal of Sociology* 83 (2): 340–363.

Nielsen, M. 2011. *Reinventing discovery: The new era of networked science*. Princeton: Princeton University Press.

O'Neil, C. 2016. Weapons of math destruction. In *Allen Lane*. London: Penguin.

Open Science Collaboration, B.A. Nosek, A.A. Aarts, C.J. Anderson, J.E. Anderson, H.B. Kappes, et al. 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251): aac4716–aac4716 ISSN 0036-8075.

Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Raghavan, P. 2014. It's time to scale the science in the social sciences. *Big Data & Society* 1(1, June): 1–4. http://bds.sagepub.com/content/1/1/2053951714532240. Accessed 9 Oct 2016.

Reality check on reproducibility. 2016. *Nature* 533(26 May): 437. https://www.nature.com/news/reality-check-on-reproducibility-1.19961. Accessed 2 Nov 2018

Rieder, G., and J. Simon. 2016. Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data and society* 3. http://bds.sagepub.com/content/3/1/2053951716649398. Accessed 9 Oct 2016.

Rodotà, S. 2014. Il mondo nella rete. Quali i diritti, quali i vincoli. Repubblica/Laterza, Rome.

Schroeck, M. et al. 2012. *Analytics: The real-world use of Big Data*, IBM Report. http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf. Accessed 9 Oct 2016.

Schroeder, R. 2014. Big Data and the brave new world of social and media research. *Big Data and society* 1(2, December): 1–11. http://bds.sagepub.com/content/1/2/2053951714563194. Accessed 9 Oct 2016.

Symons, J., and R. Alvarado 2016. Big Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society* 1–17(July–December). http://bds.sagepub.com/content/spbds/3/2/2053951716664747.full.pdf. Accessed 7 Oct 2016.

Turing A. M. 1945. Proposal for the development in the mathematical division of an automatic computing engine (ACE), Report to the Executive Committee of National Physical Laboratory 1945, reprinted in Turing 1992, Collected Works of A.M. Turing: mechanical intelligence, Ince D. C. ed. Amsterdam: North-Holland, 1–86.

———. 1947. Lecture to the London Mathematical Society on 20 February 1947. In *The essential Turing*, ed. J. Copeland, 378–394. 2004. Oxford: Clarendon Press.

Zwitter, A. 2014. Big Data Ethics. *Big Data & Society* 1(2, November): 1–6. http://bds.sagepub.com/content/1/2/2053951714559253. Accessed 9 Oct 2016.

Part IV
Cognition and Mind

Chapter 9
Telepresence and the Role of the Senses

Ingvar Tjostheim, Wolfgang Leister, and J. A. Waterworth

Abstract The telepresence experience can be evoked in a number of ways. A well-known example is a player of videogames who reports about a telepresence experience, a subjective experience of being in one place or environment, even when physically situated in another place. In this paper we set the phenomenon of telepresence into a theoretical framework. As people react subjectively to stimuli from telepresence, empirical studies can give more evidence about the phenomenon. Thus, our contribution is to bridge the theoretical with the empirical. We discuss theories of perception with an emphasis on Heidegger, Merleau-Ponty and Gibson, the role of the senses and the Spinozian belief procedure. The aim is to contribute to our understanding of this phenomenon. A telepresence-study that included the affordance concept is used to empirically study how players report sense-reactions to virtual sightseeing in two cities. We investigate and explore the interplay of the philosophical and the empirical. The findings indicate that it is not only the visual sense that plays a role in this experience, but all senses.

Keywords Affordance · Telepresence · Perceptual experience · Virtual environments · Subjective experience

9.1 Introduction

The aim of this paper is to discuss the theoretical basis for the telepresence phenomenon. We perform empirical research that takes into account the role of the senses. Both theoretical work and empirical evidence are of importance for our understanding of this phenomenon.

I. Tjostheim (✉) · W. Leister
Norsk Regnesentral, Oslo, Norway
e-mail: Ingvar.Tjostheim@nr.no

J. A. Waterworth
Umeå University, Umeå, Sweden

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_9

According to Don Ihde (1983:10), technology is a basis for an understanding both of the world and of ourselves. Technology is a broad term that refers to artifacts created by humans, such as machines, devices and components, and the methods used to create these artifacts. Telepresence is the feeling of being in a place or environment while not being physically in this environment. Telepresence can be described as a subjective experience evoked by media technology. Questions that can be asked are; why does it happen, can we identify relevant theories or theoretical contributions that we can use to discuss, analyze, and deepen our understanding of this phenomenon. Computer graphics and realistic rendering technologies play a key role in evoking telepresence. These technologies blur the lines between fiction and non-fiction. Although many developers in ICT are not very concerned with the theoretical questions, the developers seem to have a kind of understanding and feeling for how users respond to what they make. Today video games represent a major form of entertainment. As the technology has improved over the years, these games show situations and characters that are realistic and, for lack of a better term, very life-like.

von Helmholtz introduced the notion that visual perceptions are unconscious inferences, a reflex-like mechanism which is part of the formation of visual impressions. For our understanding of the phenomenon the work by von Helmholtz (1866) can be a starting point. To von Helmholtz, human perception was but indirectly related to objects, being inferred from fragmentary and often hardly relevant data signaled by the eyes. The judgment we make operates as if we were making rational inferences from sensory information – through our eyes, we necessarily perceive things as real. von Helmholtz's ideas, the type of inferences he describes, and the role of the visual sense seem relevant for telepresence.

The role of visual technologies in evoking the telepresence experience has been documented many times – there is empirical evidence for the phenomenon. Telepresence is a visual experience, and often more than that. Therefore, we will discuss telepresence as a perceptual experience and ask about the role of the senses and sense-reactions. Based on insights from the empirical work, we seek contribute to our understanding of the phenomenon, and to discuss the interplay between theory and empirical observations.

In the telepresence community, some researchers highlight the role of activities. Mel Slater et al. (2009) asks: Is telepresence better referred to as correlational presence that emphasizes the correlation between activity and sensory feedback? Furthermore, to study the phenomenon in the empirical domain of the human-technology relationship, the question is how to design a study that takes into account sense-reactions.

Some players of videogames report that their virtual experience is a veridical experience and appears similar to an experience without media technology. The telepresence researchers Lombard and Weinstein (2012: 6) give an example: they write; one of the players in the study says:

> *I completely felt that I was a part of the world and the characters and settings were all real and places I have been.*

In this context, the experience is not only described as a veridical experience, but as an actual visit to a place. Schwartz (2006: 315) has a similar observation. He quotes a player of the video game Grand Theft Auto who says:

You feel as if you're in a real town/city with other people.

In this paper, we explore theories of relevance for the telepresence phenomenon. We ask specifically about the role of the senses. If we assume that we can study telepresence in a similar manner as experiences that we have face-to-face in a non-digital environment, it seems relevant to explore the role of all senses rather than focusing mainly on the visual sense. For empirical research, the question is also how to measure the telepresence-experience. In this paper, findings from two empirical studies with video-games are used to discuss this question. Furthermore, based on a literature review, we discuss the concepts affordance and correlation presence, and align these with Merleau-Ponty's theory of perception. Finally, for the question why is telepresence perceived as a veridical experience, we draw attention to the Spinozian belief procedure, the notion that a percept is immediately believed. In decision theory this is referred to as the dual-process theories of reasoning. We are interested in the interplay between the theoretical and the empirical, and we believe this is an example of a research field where philosophy and empirical science can interact. We conclude the paper with remarks about the approaches that can be taken in future research on the telepresence phenomenon.

9.2 Perceptual Realism and the Common Sense of the Ordinary Man

Perceptual realism is the view that, in ordinary perception, one is directly aware of physical objects and events—things that exist independently of our perception of them. Most people are perceptual realists. This observation is shared by representationalists (Dretske 2003) as well as relationalists (Martin 2004). In everyday life we take the phenomenal world to be the physical world, and we treat the objects and events we perceive as if they were the objects and events themselves (Velmans 2000). Another term for this is commonsensical realism, or just common sense of the ordinary man (Putnam 1994).

People experience telepresence. A discussion of the telepresence experience as a technology-mediated experience should, therefore, include this perspective. When ordinary people are asked in an empirical study, they express their subjective feelings, their opinions based on their experiences. Although some might have a theoretical knowledge of relevance for the subject matter, it is unlikely that this is influencing how they answer questions in a survey. If I am asked whether I am having the experience of being there, the answer can be yes or no, or maybe "yes, for a moment I had this telepresence experience". We can describe this as a first person introspective judgment or report. In telepresence research there are many studies from this perspective. Many studies, particularly studies with players

of games and users of VR, seem to be more concerned with the subjective, how the players describe the experience, than the theoretical questions and theoretical contributions from other fields such as philosophy. What follows in this section is a brief review of telepresence definitions, Heidegger's and Merleau-Ponty's theory of perception, and Gibson's affordance concept.

9.3 Telepresence

Telepresence is a relatively new research field. The first journal on telepresence was inaugurated in 1992. However, since the mid-1950s, researchers have increasingly studied the telepresence phenomenon. In the 1950s the first modern VR-devices, such as the Sensorama and a number of 3D films were produced. In an article in Esprit, André Bazin (1967) entitles one section as *The Concept of Presence,* where he defines the term with regard to time and space. Bazin (1967: 96) writes:

> *Presence naturally, is defined in terms of time and space. "To be in the presence of someone" is to recognize him as existing contemporaneously with us and to note that he comes* ***within the actual range of our senses*** (our emphasis)

In the 1990s, a number of research papers were published in the telepresence journal, at conferences, and a telepresence research community was established.

The word telepresence has two parts. "Tele" refers to the Greek term *at a distance* or *far away* and is used in tele-operation and telecommunication to emphasis the remote aspect while presence refers to the here and now. Witmer and Singer (1998) define telepresence as the subjective experience of being in one place or environment, even when one is physically situated in another place. We can refer to this as telepresence as transportation. Steuer (1992:75), the author of an influential paper on telepresence, writes that:

> *when perception is mediated by a communication technology, one is forced to perceive two separate environments simultaneously: the physical environment in which one is actually present, and the environment presented via the medium. The term telepresence can be used to describe the precedence of the latter experience in favor of the former; that is, telepresence is the extent to which one feels present in the mediated environment, rather than in the immediate physical environment.*

To summarize, the attention is on the mediated environment, and the mediated environment takes precedence over the environment in which one is physically present.

Lombard and Ditton (1997) define telepresence as the perceptual illusion of non-mediation. This definition is one of the most frequently cited definitions in telepresence research. Lombard, Ditton and their colleagues, have also developed a methodology for telepresence measurement. This measurement has a sub-construct named perceptual realism concerning the five senses.

9.3.1 Telepresence and the Senses

Aristotle, in *de Anima*, writes that there are five senses. To Aristotle, vision is the primary human sense to which the others are subordinate. In the Aristotelian hierarchy of the senses, the visual sense is therefore the dominant sense (Burri et al. 2011). Also today it is common to distinguish between sight, hearing, touch, taste, and smell, the five senses (Nudds 2004). According to Fulkerson (2014) it is unlikely that we will find unified criteria for defining each of the senses. Moreover, the senses seem to be internally linked, and sometimes co-dependent. Gibson (1966) emphasized that the senses are a perceptional system. Furthermore, we can distinguish between the exteroceptive senses (such as sight and hearing) that detect objects, and properties in the external world to the body, and the interoceptive senses, which detect changes to the body (Macpherson 2011).

Telepresence can be characterized as technologically-mediated experiences, and as a medium-induced experience. The screen and the visual sense play a key role because visual media have the ability to convey non-visual aspects of perception (Merchant 2011). MacDougall (1997) suggests that the visual representation can offer pathways to the other senses. Some researchers in telepresence argue that the more the senses are stimulated, the higher the degree of presence (Sadowski 1999). When a person is experiencing a mediated or virtual reality environment *as if* the experience were non-mediated, the person is experiencing telepresence.

Some commentators maintain that telepresence can be evoked by imagining another place as well as by directly perceiving and acting in a mediated version of that place (the so-called "book problem" (Biocca 1997, Schubert 2002). Waterworth et al. (2015, p.36) write that;

> *the most relevant schism in views of presence is between those theorists who suggest that presence is evoked both by internal imagery and perceptions, and those theorists (including ourselves) who suggest that presence is evoked only in the latter case.*

Waterworth et al. (2015) see the feeling of presence in a technologically-realized place as an absorption state based around perceptual flow, essentially an equivalent experience to feeling present in the place in which the body is physically located. They suggest that imagined events and situations may also result in absorption, as in a vivid fantasy or daydream. But, people do not normally confuse what they conceive in imagination with what they perceive as the external world. They suggest that these are qualitatively different experiences with the sense of presence underlying an organism's ability to make this essential distinction. This view is compatible with the way people react bodily and perceptually as if they were physically located in a distant place, to a greater or lesser extent.

The telepresence literature presents references to theories and philosophical discussion that might help us understand the phenomenon. Two of the most influential telepresence researchers are Frank Biocca and Mel Slater, and both are concerned with such theoretical questions. For instance, Biocca suggests that presence is a sub-problem of the science of consciousness, specifically the mind-body problem. To him, virtual environments potentially alter the interaction of

the senses and motor systems with energy arrays that represent invariants of the environment such as objects, spaces, and other beings (Biocca 2001: 555).

The definition by Slater (1999) includes the following three factors: Telepresence is *a)* the sense of being there in the virtual environment; *b)* the extent to which the virtual environment becomes the dominant one, i.e., that participants respond to events in the virtual rather than the actual environment; and *c)* the extent to which participants remember having visited the place depicted in the virtual environment rather than having seen computer-generated images of it. The feeling of presence occurs when there is a successful combination of real sensory data and virtually generated sensory data or in the case of virtual reality, replacement of real sensory data (Slater et al. 2009). He argues that humans have a propensity to find correlations between their activity and internal state and their *sense perceptions* [our emphasis] of what is going on "out there." Slater (2009) is influenced by enactivism, and what is referred to as the sensorimotor approach (O'Regan and Noë 2001).

9.4 Heidegger, Merleau-Ponty and Gibson

According to Merleau-Ponty, the individual's experience rests upon the body (Low 2009). We will highlight some of Merleau-Ponty's core ideas in his theory of perception. In this context we also refer to Martin Heidegger's ready-ready-to hand concept. Merleau-Ponty does not explicitly mention Heidegger in his writings, but he has a reference to the notion *Dasein* (Matthews 2002: 5). Heidegger represents classic phenomenology. In *Being and Time* the scope is broad and goes far beyond technology. In *The Question Concerning Technology* (Heidegger 1954) he analyses the relationship with technology and modern science. He discusses technology as a means to an end and as an instrumental understanding of technology.

Techne refers to the techniques and activities that bring forth a work (poiesis), but it includes art as a process of creating. *Techne* is a mode of revealing. To Heidegger everyday activities are the starting-point and the world is *at hand* [our emphasis] in an almost-literal sense. We have a primary and pragmatic interaction with things, "*technology is a way of revealing*" (Heidegger 1954). The *ready to hand* mode is the mode of direct practical engagement in which we actually do much of our everyday living. For the relationships between Heidegger's concepts, see Fig. 9.1.

Merleau-Ponty (1962:94–95) places the body at the center of his ontology. He writes: *I am conscious of the world through the medium of my body [our emphasis].* It is from the body that I perceive the world. Merleau-Ponty does not explicitly mention Heidegger in his writings, but he has a reference to the notion *Dasein* (Matthews 2002: 5).

Merleau-Ponty argues that perception and action are linked. His research has influenced scholars in phenomenology and contemporary philosophy, but very few have used his theory and concepts as a foundation for studying experiences in virtual environments. We have only identified two researchers that refer to and discuss his work, Tripathi (2005) and Morie (2007). Tripathi argues that we are

Fig. 9.1 Heidegger and Dasein

never disembodied, not in cyberspace, not in front of a computer because we should focus on the act of experience rather on the thing being experienced. Morie (2007) emphasizes that Merleau-Ponty has paved the way for a discourse about immersive environments. She refers to Merleau-Ponty's book *The Visible and the Invisible*. Morie (2007: 107) applies Merleau-Ponty's ideas and concepts to VR and writes:

> *virtual environments are not purely imaginal; we experience them through our bodily senses, and in this way they are also real in the sense of the lived world.*

To Merleau-Ponty, things and worlds of our imagination are variations of the actual world. In *Visible and Invisible*, he states (ibid: 112):

> *it is the possible worlds and possible things that are variants and doubles of the actual world and of actual beings*.

Merleau-Ponty emphasizes that the most immediate and essential aspects of the lived dimension of space are sensory experiences. In his main work *The Phenomenology of Perception* (ibid: 239) he states:

> *by thus remaking contact with the body and with the world we shall rediscover our self, since, perceiving as we do with our body, the body is a natural self and, as it were, the subject of perception*.

The senses and perception are interrelated, and the experiences we have with our body have a meaning aspect. It is our body that actually absorbs meaning, in the form of bodily experience (ibid: 146–147). The body is both the generating and enduring aspect of experience, he writes; "*Our body is not primarily in space, it is of it.*" (ibid: 148), and existence is spatial (ibid: 342). Human subjectivity is essentially an embodied phenomenon, and there is a circular interplay between the three; body-mind-world. For instance, we have learned from our experience how to find our way around in a city. Merleau-Ponty calls this a feedback loop. Merleau-Ponty (ibid: 136) writes:

> *Cognitive life, the life of desire or perceptual life – is subtended by an intentional arc which projects round about us our past, our future, [and]our human setting.*

This intentional arc "*brings about the unity of the senses"* (ibid: 136). Merleau-Ponty explains how technology is part of the embodied experience, or how technology can be an extension of the body. His example is a blind man's use of a cane. The blind man perceives the world through his cane. This is a skill that has to be learned and a way of actively probing his environment. When he walks down the street, he is not primarily aware of the cane, instead he is aware of the curb etc. Like all other perception, it is an active communion with the world. The person's experience is created in a bodily encounter and in the reflection of this encounter. In the context of telepresence and mediated experiences, the screen and the game-console is the cane. The device becomes part of the *here-body experience* to use a term by Ihde (2002).

For a discussion of Merleau-Ponty and telepresence, there is another metaphor that should be mentioned. It is the *mirror*. Merleau-Ponty (1968) refers to Paul Schilder, an Austrian psychoanalyst and the function of a mirror. Merleau-Ponty uses the example of man with a pipe standing in front of a mirror. The mirror externalizes or extends my body, my here, in the world over there (Merleau-Ponty 1964: 129–30):

> *The mirror's phantom draws my flesh into the outer world (traîne dehors ma chair), and at the same time the invisible of my body can invest its psychic energy in the other bodies I see.*

Schilder (1935: 224) writes:

> *The experience of the sensation in the mirror is as immediate and original as the experience in the real hand.*

Some researchers in telepresence discuss the phenomenon from a theoretical point of view, but Merleau-Ponty is rarely cited. There can be many reasons why very few researchers have adopted Merleau-Ponty's theory of perception and the role of the body in studies of telepresence and presence-evoking technologies. This might indicate that insights from phenomenology and philosophy are not appreciated or understood in a field dominated by computer scientists. In our view, Merleau-Ponty's theory, his concepts and ideas, seem to be relevant not only for a theoretical understanding of the phenomenon, but for empirical work in the field of telepresence, in particular, for the choice of measurement in studies that concerns the subjective experience, that is when and why players report this *feeling of being there*.

9.4.1 *Affordance and Correlational Presence*

The perceptual psychologist James J. Gibson introduced the term affordance in his book *The Ecological Approach to Visual Perception* (1979). Since then, the affordance concept has been used in a number of disciplines other than psychology. According to Gibson (1979), an affordance is neither an objective property nor a subjective property. It is both a fact of the environment and a fact of behavior. He argues that we perceive objects as having properties of what we ought to do with them, and he attributes full normativity to affordances. Not all agree to this strong claim. Nanay (2010), for instance, holds the view that when we perceive objects, the property affords action sometimes but not always.

According to Slater et al. (2009), the feeling of presence occurs when there is a successful combination of real sensory data and virtually generated sensory data. Slater (2009) treats (tele)presence as rooted in activity, the response of people to their surroundings and their ability to actively modify those surroundings (Flach and Holden 1998; Zahorik and Jenison 1998). Slater et al. (2009) argues that presence does not demand high fidelity to physical reality, but rather that people do respond, and be able to respond, as if the sensory data were physically real, and Slater et al. 2009: 198) suggests that:

> *humans have a propensity to find correlations between their activity and internal state and their sense perceptions of what is going on out there.*

9.5 The Spinozian Belief Procedure

In his book *Ethics,* Spinoza put forward the notion that a comprehended proposition is automatically believed. This means that in the moment, we automatically accept information before being able to reject it. The proposition 49 reads (Spinoza 1982):

> *There is in the mind no volition or affirmation and negation, save that which an idea, in as much as it is an idea, involves.*

Spinoza suggested that people believe every assertion they understand, but quickly un-believe those assertions that are found to be at odds with other established facts (Gilbert 1991). Spinoza argued that to comprehend a proposition, a person implicitly accepts the proposition; only later, if the person realizes that the proposition is in conflicted with some other, he or she might change his or her mind (ebd.). Richter et al. (2009) refers to the notion of an initial acceptance of information as the dual-stage model of comprehension and validation. Stanovich (1999), Stanovich and West (2000)) labeled two types of cognitive processes, system 1 and system 2. In decision science, this is a concept often used to explain different decisions processes. There are similarities with this theory and the Spinozian belief procedure. System 1 regards intuitions that can be described as thoughts and

preferences that come to mind quickly and without much reflection (Kahneman 2002). Some formulate this belief procedure as a strong claim. Gerard (1997) writes:

> *perception is quintessentially Spinozian; a percept is immediately believed. Only in the case of rare illusions are our senses tricked into believing what is not there or in to not believing what is there.*

However, there are studies that indicate that this claim is too strong, and in some cases not an initial accept (Street and Richardson 2015). Merleau-Ponty refers to Spinoza when he discusses attention, judgement and perception, but not the Spinozian belief procedure.

Against this backdrop, we ask whether there are indications that the Spinozian belief procedure can inform our understanding of why telepresence occurs from an empirical point of view? The experience in the virtual environment can evoke what we name the telepresence experience, the feeling of being there. Telepresence is also referred to as a medium-induced experience (Steuer 1992). In the next section we present a study with video-games where the experience of a place in a VE is created.

9.6 Two Empirical Studies with Video-Games

The telepresence experience can be evoked in many ways. Schwartz (2006) argues that realism and attention to detail allow gamers to experience the game spaces as real. One of the areas in which we have seen significant technological advancement the recent years is computer graphics and computer-generated imagery. This technology seems to blur the line between fiction and non-fiction, and it plays a key role in both movies and video games. A trend in this industry is photo-realism (Leister et al. 1991). It is possible to mimic not only how humans look, but also how they behave. An example is Kara (Robinson 2012), an avatar made by the video game developer David Cage for Playstation.

The geographer Edward Relph writes that virtual places can be more or less accurate reproductions of real places and more or less convincing on their own terms (Relph 2007). In his theory, Relph (1976) proposes *vicarious sense of place*, a type of transportation to a place through imagination. Relph (1976, 2007) writes:

> *I have limited knowledge of digital virtual reality. [. . .]Nevertheless, it seems to me that mutual interaction is at work between what might be called "real" place and virtual places.*

> *It is possible to visit a place in a second hand view or vicarious way that is without actually visiting them.*

Some researchers distinguish between fantasy or imaginary places and actual places, also referred to as remote places. This is evident in the early telepresence literature. Held and Durlach (1992), Sheridan (1992), and Steuer (1992) refer to one type of the telepresence experience as telepresence *in remote places*.

Table 9.1 The profile of the participants

	Las Vegas		Los Angeles
The nationalities of the participants:	From the US	From the Netherlands	From Norway and other European countries
Age			
19–24 years old	91% (43)	41% (9)	48% (29)
25–29 years old	4% (2)	32% (7)	27% (16)
30 years or older	4% (2)	27% (6)	25% (15)
Gender			
Female	28% (13)	28% (6)	60% (36)
Male	72% (34)	72% (16)	40% (24)
N	47	22	60

9.6.1 Telepresence in Remote Places; Las Vegas and Los Angeles as Virtual Places

In many console and videogames, well-known cities are used as an urban environment and an integrated part of the narrative. In some games it is possible to explore these cities as a tourist in what is referred to as a tourist mode option. For the purpose of exploring to what extent a sightseeing experience in a virtual place evoke reactions to the senses, we chose two cities that are used in videogames: 1) Las Vegas in Project Gotham Racing 4 made by Bizarre Creations for Xbox, 2) Los Angeles in Midnight Club Los Angeles made by Rockstar for Playstation. In Table 9.1, we present the profile of the participants. The data-collection took place at three locations; a) Temple University in the US, b) Erasmus University in the Netherland and c) University of Oslo in Norway. The participants had their origins in the USA, the Netherlands, Norway, and some other European countries.

9.6.2 The Research Design of the Two Studies with the Cities Las Vegas and Los Angeles

Both cities are presented in a photorealistic manner in the games. The visuals from both games were used unchanged, but the sound was substituted with an audio-guide for tourists in order to create a sightseeing experience. The audio-guide the "*Hollywood Audio Tour*" by the company *Tourcaster* was combined with the videogame. In the studies, the participants were not given any information about the game itself, just the name of the city.

For the Las Vegas study a between-group design was chosen and the participants were randomly assigned to two groups. For both groups the task was to take part in sightseeing in Las Vegas. The participants all listened to a guide and looked at the buildings along *the Strip* on a big screen. The sightseeing tour lasted 7 min.

For the first group a photo-mode setting was used. Photo-mode is similar to a recorded slideshow that present pictures one by on while the guide is talking about the buildings and the history of the city, the hotels and casinos. The other group had a similar presentation, but in motion-mode. The motion-mode is default for players of the game. However, by comparing this to a photo-mode, a slideshow of pictures, the motion-effect can be revealed.

For the Los Angeles-study the sightseeing was a live event in the sense that the visual of the game was used without any adaptation. The introduction was:

> *You are now going to do sightseeing in LA on the screen in front of you*." And, "*I, the interviewer will be a co-guide and tell you when to move forward, when to stop and listen to the guide*.

All participants started on Vine Street with a view of the Capital Records Tower, and continued into Hollywood Boulevard. The virtual sightseeing tour lasted for approximately 15 min.

There are six measuring instruments that are commonly used in telepresence research. The purpose is to capture the subjective experience of the player (Nunez 2007). These six instruments are; the Slater et al. (1994)), the Presence Questionnaire (PQ) (Witmer and Singer 1998), the Igroup Presence Questionnaire (IPQ) (Schubert et al. 2001), the Independent Television Commission's Sense of Presence Inventory (Lessiter et al. 2001), the MEC Special Presence (MEC-SPQ, Vorderer et al. 2004), and Temple Presence Inventory (TPI) (Lombard et al. 2000; Lombard et al. 2011).

For this study, the TPI was chosen because it contains the sub-construct named perceptual realism about the five senses; sight, smell, touch, sound and taste.

We hypothesized, based on our theoretical discussion that it is when the senses are evoked that a telepresence experience occurs, or a stronger telepresence experience occurs. In order to study the correlation between activities and sensory feedback, the affordance concept was chosen.

9.6.3 The Key Findings

First we report the mean scores, see Table 9.2.

The higher scores for the motion mode indicate that moving pictures have a stronger telepresence effect on the participants than still photographs. This is in accordance with both theory and findings in other empirical studies (Yoon et al. 2008; Ozok and Komlodi 2009).

For the Los Angeles study, all participants did an interactive sightseeing. With the exception of smell, it seems that, compared to the two alternatives, the interactive sightseeing experience evokes a stronger telepresence experience, that is, a higher score on four of five senses. We counted the number of participants that answered agree on the senses touch, look and sound, that is four on one of them and five or higher on the other two. We can describe those belonging to this group as

Table 9.2 The Las Vegas and the Los Angeles studies and perceptual realism

	Las Vegas	Las Vegas	Los Angeles
	Sightseeing in photo-mode	Sightseeing in motion-mode	Interactive sightseeing in game-mode
	Mean	Mean	Mean
Perceptual realism, 7-point scale (1 = fully disagree, 7 = fully agree)	2.85	3.40	4.27
Overall how much did touching the things and people in the city you saw feel like it would if you had experienced them directly?			
How much did the heat or coolness (temperature) of the city you saw feel like it would if you had experienced it directly?	2.74	3.09	3.23
Overall, how much did the things and people in the city you saw smell like they would if you had experienced them directly?	1.97	2.54	2.33
Overall, how much did the things and people in the city you saw look like they would if you had experienced them directly?	3.09	3.97	4.60
Overall, how much did the things and people in the city you saw sound like they would if you had experienced them directly?	2.62	3.71	3.60

"senses evoked". A mean score around four can be interpreted as neither negative nor positive. In the Las Vegas study, in the motion-mode, 12 of the 35 participants reported a positive sense-reaction. In the Los Angeles study 18 of the 60 participants reported that the virtual sightseeing evoked a sense-reaction, see Table 9.3.

Secondly, it is pertinent to ask how are the 18 (30%) that have the telepresence-experience different from the 35 that did not report they had this feeling of

Table 9.3 Affordances in a virtual environment

being there? Mel Slater claims that humans have a propensity to find correlations between their activity and internal state and their sense perceptions. This is the key argument for the concept named correlational presence. Is there empirical evidence for this claim? Correlational presence and affordance are closely related. One of the purposes of the empirical study was to use these concepts together with measurements from telepresence.

Table 9.3 shows a pattern. The findings indicate that the participants in the third group had a telepresence experience. The participants in the senses-not-evoked group were different. The numbers indicate that most participants in this group did not have a telepresence experience.

9.7 Discussion

In most cases the person that has *the feeling of being there* in a virtual place knows that a medium is involved. We do not discuss this question any further, but we agree with Floridi (2005) who argues that we should not define something as complex as presence by what it is not and by the failure of someone not to notice something.

Many telepresence studies only report whether or not the feeling of being there is experienced in the moment or immediately after. We designed a study with a sightseeing experience. We have documented that some of the participants reported that it was an experience of the place, a feeling of being in the actual city. The participants reported this immediately after the sightseeing had ended.

Baruch Spinoza rejected the mind-body dualism of Descartes. One of his propositions concerns how we react and make judgment when receiving information.

Heinemann (1941) with references to the empiricist school founded by Philinos of Kos in Alexandria distinguishes between three sorts of experiences. These are:

> *immediate experience, mediated experience (that is observation made by others before us), and analogous experience (thus in case of illness which has not been observed it may be useful to compare similar cases).*

Heinemann, in his discussion on types of experiences, refers J. A. H. Murray, the *A New Dictionary on Historical Principles* (Oxford 1817). Murray distinguishes between to have an experience of, to learn by experience and to try something, a tentative experience. Regarding to have an experience, the first of these three Heinemann writes (1941: 570):

> *(i) To have, experience of; to meet with; to feel; to suffer; to undergo. We could call this an immediate experience; it covers what we immediately feel or undergo during the course of our life.*

The immediate experience corresponds most often to how we use the word in this paper, and for instances in the phrase "*an experience in a VE*." We emphasize the present tense, the experiencing.

There can be many answers to the question why this feeling of being there occurs although the person knows that it is a media-induced experience. This question can be investigated with different lenses, and within an interdisciplinary context. In this paper we have drawn attention to some of the theories and ideas from phenomenology and philosophy, theories that can be used to reflect on what telepresence is and why it happens. And we have briefly discussed Merleau-Ponty and some of his thoughts and the Spinozian belief procedure.

Experiences also include perceiving through the senses, as well as feeling and doing. Logue (2009) defines perceptual experience as experience associated with sense modalities (vision, hearing touch, smell and taste) in virtue of which it appears to one that one's environment is a certain way. He emphasis is on the word *of*. He posits that a perceptual experience is a matter of a certain sort of relation obtaining between the subject of the experience and what the experience is of, that is the object of the experience.

To design studies and investigate the interplay of the empirical and the philosophical is not an easy task. Technology plays a key role in our society and research that can contribute to theoretical discussion of experiences in which technology plays a key role should be encouraged.

The telepresence researcher Sheridan (1992) considers the extent of sensory information provided by media technology to be a mjor factor contributing to telepresence. According to Mingers (2001), the success of VR will depend on the extent to which it can mimic a response to all the nervous system's sensory modalities. Not all will agree, but in the history of (console and PC) games, there are examples of games with simple graphics that can create a sense of presence. It is, however, pertinent to study the role of the senses with regard to telepresence and experiences in virtual environments. With video-games and video-game technologies, there are many opportunities for empirical studies, to test hypotheses about the role of the senses.

We have based our studies on virtual environments of cities from two video-games that are made by professional game-developers. The results from the two empirical studies indicate that the experience in the virtual environments evoked a bodily reaction for some of the participants, but not for all of them. The main contribution of this study is to draw attention to the need for a theoretical discussion about telepresence that includes phenomenology and theories of perception. For empirical work, we have given an example with the affordance concept and how sense-reactions can be measured in telepresence studies.

McLuhan (1964) stated that media are extensions of the senses. Steuer (1992) had the vision that media technologies become more and more vivid. Thus, it is possible that we will, in the future, experience that systems will be capable to pass a version of the imitation game (Turing 1950) that we can refer to as a "perceptual Turing test". We are not there yet, but theories on the belief procedures, why a percept is believed, and the role of the body should be in our inquiries and analysis.

The affordance concept can be operationalized and used in empirical studies. The development of decision theories is often based on empirical work. Insights from this field seem relevant for a discussion of the telepresence phenomenon. For future research, we should ask; are there good alternatives to the survey-based approach, that is, to ask the person to report to what extent the person has a telepresence experience? In addition to asking participants, that is, use introspective methods, we believe that electroencephalography (EEG), biosensors, and similar technology will play and role in telepresence research in the future. Such technologies are suitable to monitor sense-reactions in the moment. There is already research along this path. An example is the neurophysiological study by Baumgartner et al. (2006) on electro-encephalography and spatial presence, a functional magnetic resonance imaging (fMRI) VR-study by Hoffman et al. (2003), and the study by Clemente et al. (2013) on telepresence and the activity of the right insula in the brain. This leads to an intricate question: how should we interpret this type of data without asking the person about the subjective experience?

References

Baumgartner, T., L. Valko, M. Esslen, and L. Jäncke. 2006. Neural correlate of spatial presence in an arousing and noninteractive virtual reality: An EEG and psychophysiology study. *Cyberpsychology & Behavior* 9 (1): 30–45.

Bazin, A. 1967. *What Is Cinema?* Trans. H. Gray. Los Angeles: University of California Press. (Original work published 1951).

Biocca, F. 1997. The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer Mediated Communication* 3(2). Available at: http://jcmc.indiana.edu/vol3/issue4.

———. 2001. Inserting the presence of mind into a philosophy of presence: A response to Sheridan and Mantovani and Riva. *Presence: Teleoperators and Virtual Environments* 10 (5): 546–556.

Burri, R.V., C. Schubert, and J. Strübing. 2011. The five senses of science. *Science, Technology & Innovation Studies* 7 (1): 1–3.

Clemente, M., A.J. Rodríguez, B. Rey, and M. Alcañiz. 2013. Measuring presence during the navigation in a virtual environment using EEG. In *Annual review of cybertherapy and*

telemedicine 2013: Positive technology and health engagement for healthy living and active ageing, ed. B.K. Wiederhold and G. Riva, vol. 191, 136–140. Amsterdam: IOS Press.
Dretske, F. 2003. Experience as representations. *Philosophical Issues, Philosophy of Mind* 13: 67–82.
Flach, J.M., and J.G. Holden. 1998. The reality of experience. *Presence, Teleoperators, and Virtual Environments* 7: 90–95.
Floridi, L. 2005. The philosophy of presence: From epistemic failure to successful observation. *Presence: Teleoperators and Virtual Environments* 14 (6): 656–667.
Fulkerson, M. 2014. Rethinking the senses and their interactions: The case for sensory pluralism. *Frontiers in Psychology* 5: 1426.
Gerard, H.B. 1997. Psychic reality and unconscious belief: A reconsideration. *International Journal of Psychoanalysis* 78: 327–334.
Gibson, J.J. 1966. *The senses considered as perceptual systems*. Westport: Greenwood Press.
Gilbert, D.T. 1991. How mental systems believe. *American Psychologist* 46 (2): 107–119.
Heidegger, M. 1954. *The question concerning technology, and other essays*, 1977. New York: Harper & Row.
Heinemann, F.H. 1941. The analysis of 'Experience'. *The Philosophical Review* 50 (6): 561–584.
Held, R.M., and N.I. Durlach. 1992. Telepresence. *Presence* 1 (1): 109–112.
Hoffman, H.G., T. Richards, B. Coda, A. Richards, and S.R. Sharar. 2003. The illusion of presence in immersive virtual reality during an fMRI brain scan. *Cyberpsychology & Behavior* 6 (2): 127–131.
Ihde, D. 1983. *Existential technics*. Albany: SUNY Press.
———. 2002. *Bodies in technology*. Minneapolis: University of Minnesota Press.
Kahneman, D. 2002. *Maps of bounded rationality* (No. 2002–4). Nobel Prize Committee.
Leister, W., H. Müller, and A. Stößer. 1991. *Fotorealistische Computeranimation*. Springer. ISBN 3-540-53234-X, in German.
Lessiter, J., J. Freeman, E. Keogh, and J. Davidoff. 2001. A cross-media presence questionnaire: The ITC-sense of presence inventory. *Presence: Teleoperators and Virtual Environments* 10: 282–298.
Logue, H. 2009. *Perceptual experience: relations and representations.* Doctoral dissertation, Massachusetts Institute of Technology.
Lombard, M., and T.B. Ditton 1997. At the heart of it all: The concept of presence. Journal of Computer-Mediated Communication 3(2).
Lombard, M., T.B. Ditton, D. Crane, B. Davis, G. Gil-Egui, and K. Horvath. 2000. Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. In *Proceedings of the third international workshop on presence*, ed. W. IJsselsteijn, J. Freeman, and H. de Ridder.
Lombard, M., T. B. Ditton, and L. Weinstein. 2011. Measuring telepresence: The validity of the Temple Presence Inventory (TPI) in a gaming context. *Fourteenth International Workshop on Presence* (ISPR 2011), Edinburgh, Scotland.
Lombard, M., and Weinstein, L. 2012. What are telepresence experiences like in the real world? A qualitative survey. In Proceedings of the 15th international workshop on presence (ISPR'14). https://ispr.info/presence-conferences/previous-conferences/ispr-2014/. Accessed 16 Oct 2018.
Low, D. 2009. The body of Merleau – Ponty's work as a developing whole. *International Philosophical Quarterly* 49 (2): 207–227.
MacDougall, D. 1997. The visual in anthropology. In *Rethinking visual anthropology*, ed. M. Banks and H. Morphy. London: Routledge.
Macpherson, F. 2011. Individuating the senses. In *The senses*, 3–43. New York: Oxford University Press.
Martin, M.G.F. 2004. The limits of self-awareness. *Philosophical Studies* 120 (1–3): 37–89.
Matthews, E. 2002. *The philosophy of Merleau-Ponty*. Stocksfield: Acumen Publishing.
McLuhan, M. 1964. *Understanding media: The extensions of man*. New York: McGraw Hill.
Merchant, S. 2011. The body and the senses: Visual methods, videography and the submarine sensorium. *Body & Society* 17 (1): 53–72.

Merleau-Ponty, M. 1962. *Phenomenology of Perception.* Trans. C. Smith. New York: Routledge (Original work published in 1945).

———. (1968). *The Visible and the Invisible.* Trans. A. Lingis. Evanston: Northwestern University Press.

Mingers, J. 2001. Embodying information systems: The contribution of phenomenology. *Information and Organization* 11: 103–128.

Morie, J.F. 2007. *Meaning and emplacement in expressive immersive virtual environments*. Doctoral dissertation, University of East London.

Nanay, B. 2010. Action-oriented perception. *European Journal of Philosophy* 20 (3): 430–446.

Nudds, M. 2004. The significance of the senses. *Proceedings of the Aristotelian Society* 104 (1): 31–51.

Nunez, D. 2007. A capacity limited, cognitive constructionist model of virtual presence. Unpublished PhD thesis, Department of Computer Science, University of Cape Town, South Africa.

O'Regan, K.J., and A. Noë. 2001. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24: 939–973 discussion 973-1031.

Ozok, A.A., and A. Komlodi. 2009. Better in 3D? An empirical investigation of user satisfaction and preferences concerning two-dimensional and three-dimensional product representations in business-to-consumer e-commerce. *International Journal of Human-Computer Interaction* 25 (4): 243–281.

Putnam, H. 1994. Sense, nonsense and the senses: An inquiry into the powers of the human mind. *The Journal of Philosophy* 91 (9): 445–517.

Relph, E. 1976. *Place and placelessnes*. London: Pion.

———. 2007. Spirit of place and sense of place in virtual realities. *Techne. Research in Philosophy and Technology. Special Issue: Real and Virtual Places* 10 (3): 17–24.

Richter, T., S. Schroeder, and B. Wöhrmann. 2009. You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology* 96 (3): 538–558.

Robinson, M. 2012. Introducing Quantic Dream's Kara. In Eurogamer.net, March 7, 2012. Available at: www.eurogamer.net/articles/2012-03-07-introducing-quantic-dreams-kara

Sadowski, W. 1999. Special report: Utilization of olfactory stimulation in virtual environments. *VR News* 8 (4): 18–21.

Schilder, P. 1935. *The image and appearance of the human body*. London/New York: International University Press.

Schubert, T. 2002. Five theses on the book problem: Presence in books, film, and VR. In *5th annual international workshop presence 2002*, ed. F. Gouveia, 53–58. Porto: Universidare Fernando Pessoa.

Schubert, T., F. Friedmann, and H. Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments* 10: 266–281.

Schwartz, L. 2006. Fantasy, realism, and the other in recent video games. *Space and Culture* 9 (3): 313–325.

Sheridan, T.B. 1992. Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments* 1 (1): 120–125.

Slater, M. 1999. Measuring presence: A response to the Witmer and Singer Presence Questionnaire. *Presence: Teleoperators and Virtual Environments* 8 (5): 560–565.

Slater, M. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society*. B, 364: 3549–3557.

Slater, M., M. Usoh, and A. Steed. 1994. Depth of presence in virtual environments. *Presence* 3 (2): 130–144.

Slater, M., B. Lotto, M.M. Arnold, and M.V. Sanchez-Vives. 2009. How we experience immersive virtual environments: The concept of presence and its measurement. *Anuario de Psicología* 40: 193–210.

Spinoza, B. 1982. *The Ethics and Selected Letters*. Ed. S. Feldman, and Trans. S. Shirley. Indianapolis: Hackett. (Original work published 1677). Translated from the Latin by R.H.M. Elwes (1883), *MTSU philosophy WebWorks* Hypertext Edition 1997.

Stanovich, K.E. 1999. *Who is rational? Studies of individual differences in reasoning*. Mahwah: Erlbaum.

Stanovich, K.E., and R.F. West. 2000. Advancing the rationality debate. *Behavioral and Brain Sciences* 23 (05): 701–717.

Steuer, J. 1992. Defining virtual reality: Dimensions determining telepresence. *Journal of Communication* 42 (4): 73–92.

Street, C.N., and D.C. Richardson. 2015. Descartes versus Spinoza: Truth, uncertainty, and bias. *Social Cognition* 33: 1–12.

Tripathi, A.K. 2005. Computers and the embodied nature of communication: Merleau-Ponty's new ontology of embodiment. *ACM Ubiquity* 6 (44): 1–17.

Turing, A.M. 1950, October. Computing machinery and intelligence. *Mind* 54 (236): 433–460.

Velmans, M. 2000. *Understanding consciousness*. London/Philadelphia: Routledge.

von Helmholtz, H. 1866. Concerning the perceptions in general. In *Treatise on physiological optics,* vol. III, 3rd ed. Trans. J.P.C. Southall 1925 Opt. Soc. Am. Section 26, New York: Dover, 1962.

Vorderer, P., W. Wirth, F. R. Gouveia, F. Biocca, T. Saari, et al. 2004. MEC Spatial Presence Questionnaire (MECSPQ). Report to the European Community, project presence: MEC (IST-2001-31661).

Waterworth, J.A., E.L. Waterworth, G. Riva, and F. Mantovani. 2015. Presence: Form, content and consciousness. In *Immersed in media: Telepresence theory, Measurement & Technology*, ed. M. Lombard, J. Freeman, W. IJsselsteijn, and R.J. Schaevitz, 35–58. Berlin: Springer.

Witmer, B.G., and M.J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments* 7: 225–240.

Yoon, S., J. Laffey, and H. Oh. 2008. Understanding usability and user experience of web-based 3D graphics technology. *International Journal of Human-Computer Interaction* 24 (3): 288–306.

Zahorik, P., and R.L. Jenison. 1998. Presence as being-in-the-world. *Presence: Teleoperators and Virtual Environments* 7: 78–89.

Chapter 10
Ontologies, Mental Disorders and Prototypes

Maria Cristina Amoretti, Marcello Frixione, Antonio Lieto, and Greta Adamo

Abstract As it emerged from philosophical analyses and cognitive research, most concepts exhibit typicality effects, and resist to the efforts of defining them in terms of necessary and sufficient conditions. This holds also in the case of many medical concepts. This is a problem for the design of computer science ontologies, since knowledge representation formalisms commonly adopted in this field (such as, in the first place, the Web Ontology Language – OWL) do not allow for the representation of concepts in terms of typical traits. However, the need of representing concepts in terms of typical traits concerns almost every domain of real world knowledge, including medical domains. In particular, in this article we take into account the domain of mental disorders, starting from the DSM-5 descriptions of some specific mental disorders. On this respect, we favor a hybrid approach to the representation of psychiatric concepts, in which ontology oriented formalisms are combined to a geometric representation of knowledge based on conceptual spaces.

Keywords Representation of concepts · Formal ontologies · Conceptual spaces · Medical ontologies · Mental disorders · DSM-5 · Prototypes · Exemplars

M. C. Amoretti (✉) · M. Frixione
DAFIST, Philosophy Section, University of Genoa, Genoa, Italy

Center for the Philosophy of Health and Disease, Genoa, Italy
e-mail: cristina.amoretti@unige.it

A. Lieto
Department of Computer Science, University of Turin, Turin, Italy

ICAR-CNR, Palermo, Italy

G. Adamo
Bruno Kessler Foundation, Trento, Italy

DIBRIS, University of Genoa, Genoa, Italy

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_10

10.1 Introduction

As it emerged from philosophical analyses and cognitive research, most concepts exhibit typicality effects, and resist to the efforts of defining them in terms of necessary and sufficient conditions. This holds also in the case of many medical concepts. This is a problem for the design of computer science ontologies, since knowledge representation formalisms commonly adopted in this field (such as, in the first place, the Web Ontology Language – OWL) do not allow for the representation of concepts in terms of typical traits. However, the need of representing concepts in terms of typical traits concerns almost every domain of real world knowledge, including medical domains. In this article we take into account the domain of mental disorders, starting from the DSM-5 descriptions of some specific mental disorders. On this respect, we favor a hybrid approach to the representation of psychiatric concepts, in which ontology oriented formalisms are combined to a geometric representation of knowledge based on conceptual spaces.

In Sect. 10.2. we shall expose some problems faced by the classical theory of concepts, according to which concepts can be defined through necessary and sufficient conditions. In particular, we shall examine the important issues raised by conceptual "typicality", which concerns both common-sense and medical concepts, focusing on the general concept of MENTAL DISORDER and the various concepts of individual mental disorders as described by DSM-5. In Sect. 10.3. we shall briefly summarize the most common way to deal with the problem of concept representation, which received a great deal of attention within the field of artificial intelligence (AI), due to its relevance for semantic technologies and for the development of formal ontologies. We shall maintain that the most representative formalisms currently adopted for the development of formal ontologies, known as description logics (DLs), are unfortunately unable to represent concepts in prototypical terms. In Sect. 10.4. we shall describe an ontology we specifically build to represent the general concept of MENTAL DISORDER and (most of) the various concepts of individual mental disorders. Despite the fact that there already are formal ontologies dealing with mental disorders, we decided to develop a new one trying to overcome some of their potential limitations. As our formal ontology, despite being more DSM-5 compliant than others, is still unable to handle typicality effects, in Sect. 10.5. we shall propose a hybrid approach combining a "classical" component (in which concepts are represented in terms of necessary and/or sufficient conditions) with a "typicality-oriented" component, allowing both prototype and exemplar-based representations.

10.2 Representing Concepts: Some Problems Raised by Medicine

In philosophy and cognitive sciences, different theories about the nature of concepts have been proposed. According to the traditional view, known as "classical", concepts can be simply defined in terms of sets of necessary and sufficient conditions. This theory was dominant since the times of Aristotle until the mid'70s of the last century, when the philosophical analyses by Ludwig Wittgenstein (Wittgenstein 1953) and the experimental results obtained by Eleaonor Rosch (Rosch 1975; Rosch and Mervis 1975) showed that, for most of the common-sense concepts, this position does not hold since conceptual structures are mainly characterized by "typical" category membership cues and thus suggested that are organized in human mind in terms of prototypes. Since then, different positions and theories on the nature of concepts have been proposed in order to explain the aspects concerning conceptual "typicality". Usually, they are grouped in three main classes, namely: prototype views, exemplar views and theory-theories (see e.g. Murphy 2002; Machery 2010). All of them are assumed to account for (some aspects of) prototypical effects in conceptualization.

According to the prototype view (introduced by Rosch), knowledge about categories is stored in terms of prototypes, i.e. in terms of some representation of the "best" instances of the category. For example, the concept CAT should coincide with a representation of a prototypical cat. In the simpler versions of this approach, prototypes are represented as (possibly weighted) lists of features.

According to the exemplar view, a given category is mentally represented as a set of specific exemplars explicitly stored within memory: the mental representation of the concept CAT is the set of the representations of (some of) the cats we encountered during our lifetime.

Theory-theory approaches adopt some form of holistic point of view about concepts. According to some versions of the theory-theories, concepts are analogous to theoretical terms in a scientific theory. For example, the concept CAT is individuated by the role it plays in our mental theory of zoology. In other version of the approach, concepts themselves are identified with micro-theories of some sort. For example, the concept CAT should be identified with a mentally represented micro-theory about cats.

These approaches turned out to be not mutually exclusive. Rather, they seem to succeed in explaining different classes of cognitive phenomena, and many researchers hold that all of them are needed to explain psychological data (see again Murphy 2002; Machery 2009).

The case of some medical concepts, such as the general concept of DISEASE and the various individual disease concepts (such as PNEUMONIA, BREAST CANCER, SCHIZOPHRENIA, BORDERLINE PERSONALITY DISORDER, and so on) show the same "problems" presented by most common-sense concepts, as they can hardly be represented in terms of individually necessary and jointly sufficient conditions. Faced with the issues raised by the many attempts to find a

traditional definition for the general concept of DISEASE (Amoretti 2015), some philosophers of medicine have thus proposed to regard the concept of DISEASE and those of individual diseases as non-classical ones. In this vein, on the grounds of the great variability among individual diseases, new theories based on family resemblances, prototypes or exemplars have been proposed (see e.g. Sadegh-Zadeh 2000, 2008, 2011; Lilienfeld and Marino 1995, 1999; Pickering 2013, 2016; McNally 2011).

In the case of family resemblances, there is no common feature that all individual diseases must have, but any two of them should share at least one feature. In the case of prototypes, there is a set of properties that represents the best instance of the disease category, that is an ideal and abstract construction of the general concept of disease, the prototype, to which any individual disease must approximate to some degree, sharing with it a goodly number of properties. In the case of exemplars, some individual diseases are regarded as particularly relevant, as the exemplars of the disease category, and thus all other diseases must exhibit a goodly number of their specific features.

These views are obviously different: embracing the family resemblances theory implies that there is no specific set of properties, determined by the prototype or the exemplars, that individual diseases must meet to some degree; the prototype is an abstract construction that doesn't need to correspond to any individual disease, while the exemplars are concrete members of the category. Nevertheless they are often conflated or muddled in the relevant literature.

For example, McNally (2011, p. 212) refers to Wittgenstein saying that "Examples of most useful concepts bear only a family resemblance to one another. Most have some overlapping attributes without sharing an essence present in every case"; but clearly he has in mind the prototype view, as he continues specifying that "The more attributes a given case has, the better an example it is of the concept". A similar confusion is made by Cooper (2007, p. 41), who mentions family resemblances saying that "While there need not be any one feature that all family members possess, any two members will be similar in a variety of ways"; however, she unpacks this idea through the exemplar view: "whether a condition counts as a mental disorder depends on its degree of resemblance to prototypical cases, such as schizophrenia and psychotic depression. Conditions that are sufficiently like these central cases get counted as disorders". Again, Sadegh-Zadeh (2008, p. 119) seems to conflates prototypes and exemplars claiming that "A concept determines a category [...] by exhibiting the relational structure of the category that is characterized by best examples, called prototypes, such that other category resemble them to different extents".

The above confusions can be partially explained by the fact that all three views offer a plausible way to deal with conceptual "typicality", that is the evidence that some instances of the general category of disease, namely some individual diseases, are regarded as more representative than others. Moreover, all three views agree that there is no set of properties shared by all and only individual diseases: no specific property is individually necessary and no fixed number of them is sufficient to characterize the general concept of DISEASE. On the contrary, overall

similarities among different set of properties should encompass the absence of any particular shared property – such as, as it is often claimed, dysfunction (Boorse 1976; Wakefield 1992, 1999).

Many scholars adopting one of the above strategies do not attempt to better explicate the similarities relationship among individual diseases (Lilienfeld and Marino 1995, 1999); others think that fuzzy logic is the best, and possibly the only, way do the job (Seising and Tabacchi 2013; Sadegh-Zadeh 2000, 2008, 2011) – but, of course, some important alternatives to represent non-classical concepts have been proposed in the general literature, especially because fuzzy-logic faces some unavoidable difficulties in handling typicality (on this aspect see Frixione and Lieto 2014a).

As sketched above, approaches based on family resemblances, prototypes, and exemplars have been used to characterize the general concept of DISEASE, but they seem particularly suited to handle the general concept of MENTAL DISORDER (Lilienfeld and Marino 1995, 1999) as well as the various concepts of individual mental disorders. This more restricted class of medical concepts will be the focus of our present work.

The DSM (the Diagnostic and Statistic Manual of Mental Disorders), which is published by the American Psychiatric Association and represents a sort of "bible" for psychiatrists and scholars within the field of mental pathology, has in fact a merely descriptive approach: it rarely incorporates theoretical information regarding the causes of individual mental disorders, and classifies them using a list of operational diagnostic criteria. As a consequence, and somehow differently to what usually happens with individual somatic diseases included in ICD (the International Classification of Disease), individual mental disorders are typically identified not by their etiology or underlying pathological cause (a few exceptions being, for example, the different types of neurocognitive disorders), but through their syndromes, that is through a catalogue of their characterizing symptoms and signs; in most cases none of them is individually necessary and no fixed number of them is sufficient to determine membership to a certain individual disorder category. Moreover, in most cases these syndromes are not supposed to be reified, as to correspond to some kind of entity or mechanism (such as an underlying dysfunction).

Let's see, for instance, an oversimplified version of the diagnostic criteria for schizophrenia and borderline personality disorder given by the DSM-5. Criterion A for schizophrenia states:

> (A) *Two (or more)* of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated). At least one of these must be (1), (2), or (3):
>
> 1. Delusions.
> 2. Hallucinations.
> 3. Disorganized speech (e.g., frequent derailment or incoherence).
> 4. Grossly disorganized or catatonic behavior.
> 5. Negative symptoms (i.e., diminished emotional expression or avolition) (American Psychiatric Association 2013, p. 99, our italics).

Similarly, but even more explicatory, borderline personality disorder is characterized as follows:

> A pervasive pattern of instability of interpersonal relationships, self-image, and affects, and marked impulsivity, beginning by early adulthood and present in a variety of contexts, as indicated by *five (or more)* of the following:
>
> 1. Frantic efforts to avoid real or imagined abandonment. [. . .]
> 2. A pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation.
> 3. Identity disturbance: markedly and persistently unstable self-image or sense of self.
> 4. Impulsivity in at least two areas that are potentially self-damaging (e.g., spending, sex, substance abuse, reckless driving, binge eating). [. . .]
> 5. Recurrent suicidal behavior, gestures, or threats, or self-mutilating behavior.
> 6. Affective instability due to a marked reactivity of mood (e.g., intense episodic dysphoria, irritability, or anxiety usually lasting a few hours and only rarely more than a few days).
> 7. Chronic feelings of emptiness.
> 8. Inappropriate, intense anger or difficulty controlling anger (e.g., frequent displays of temper, constant anger, recurrent physical fights).
> 9. Transient, stress-related paranoid ideation or severe dissociative symptoms (American Psychiatric Association 2013, p. 663, our italics).

It is easy to see that there are many different ways to meet the requirements of schizophrenia or bipolar personality disorders stated above, and it is of course possible that members of these categories have no characteristics in common. For example, Galatzer-Levy and Bryant (2013) recently calculated that there are 636,120 ways to meet the requirements of the concept of post-traumatic stress disorder.

The operational criteria, introduced in DSM-III (1982), were meant to replace what psychiatrists dub as "prototypes", that is short descriptions of paradigmatic cases that would serve as standards of comparison to evaluate and diagnose any single patient. Here, as an example, the category of schizophrenic reactions according to DSM-I (1952):

> It represents a group of psychotic reactions characterized by fundamental disturbances in reality relationships and concept formations, with affective, behavioral, and intellectual disturbances in varying degrees and mixtures. The disorders are marked by strong tendency to retreat from reality, by emotional disharmony, unpredictable disturbances in stream of thought, regressive behavior, and in some, by a tendency to deterioration, (American Psychiatric Association 1952, p. 26).

Even if the operational structure of DSM-5 coincides neither with the prototype nor the exemplar views as they are developed by cognitive psychologists, it may still suggest to incorporate some features of these approaches in the representations of the various concepts of individual mental disorders (such as, SCHIZOPHRENIA, BORDERLINE PERSONALITY DISORDER, MAJOR DEPRESSION, etc.) as well as the general concept of MENTAL DISORDER, as like non-classical concepts they cannot be possibly defined through necessary and sufficient conditions, and clearly exhibit prototypical effects.

In order to address this problem from a computational perspective, we have analyzed the field of logic-oriented knowledge representation systems and, in particular, the class of formalisms known as formal ontologies. We provide below a brief overview of this class of systems by showing that, also in this artificial context, we face the problem of representing typical or "non-classical" information of medical concepts.

10.3 Formal Ontologies and Common-Sense Representations

In the last decades the problem of concept representation received a great deal of attention within the field of artificial intelligence (AI), and in particular in knowledge representation, due to its relevance for semantic technologies and for the development of formal ontologies.

In the AI tradition, an ontology is "an engineering artifact, constituted by a specific vocabulary used to describe a certain reality, plus a set of explicit (axiomatic) assumptions regarding the intended meaning of the vocabulary words" (Guarino 1998). The representation languages adopted for the development of formal ontologies stemmed from the tradition of so-called structured inheritance semantic networks – the first system in this line of research was KL-ONE (Brachman and Schmolze 1985). These formalisms are known today as description logics (DLs), and the main formal ontological languages such as OWL and OWL 2 belong to this class. The main constructs of such languages are concepts (or classes), roles (or properties), and individuals.

DLs are logical systems (usually, they are subsets of first order predicate calculus). They can perform a series of important automatic inferences, such as categorization (the process of attributing a specific individual as a member to a class), classification (the process through which new class-subclass relations are inferred) and consistency checking (the process of testing the logical coherence of a given ontology).

As logical systems, DLs have a model theoretic, Tarskian style semantics associated to them (Horrocks et al. 2003). This fact is a symptom of a problem: Tarskian semantics is fully compositional, and typicality effects are hard to accommodate with compositionality (Fodor 1981). Indeed, DLs do not allow the representation of concepts in prototypical terms (on this aspect see Frixione and Lieto 2012). DLs allow the representation of concepts exclusively in terms of sets of necessary and/or sufficient conditions. This is a severe drawback from the standpoint of the representation of many classes of concepts.

In particular, this strong bias towards the representation of concepts in terms of necessary and/or sufficient conditions alone is also a problem in the field of medical ontologies. Most of them, indeed, (including SNOMED, the largest biomedical ontology currently available: http://www.ihtsdo.org/snomed-ct) are conditioned by the adoption of formalisms that do not allow to represent concepts in typical terms.

This possibility should be of crucial importance for representing both such general concepts as DISEASE or MENTAL DISORDER, and the concepts of individual diseases and mental disorders. Consider for example the concepts of individual mental disorders. In DSM-5 they are characterized in terms of syndromes and operational criteria. However, at the level of specific mental disorders, it is often impossible to individuate sets of symptoms and criteria that are individually necessary and/or jointly sufficient to determine membership.

10.4 Some Preliminaries of a Case Study: The Schizophrenia Spectrum

As a preliminary step to prove our hypothesis that the general concept of MENTAL DISORDER and (most of) the various concepts of individual mental disorders – as they are currently described and categorized by DSM-5 – should be treated as non-classical ones is preliminary tested by developing an ontology based on the OWL-DL (Ontology Web Language Description Logic) dialect. Some important remarks must be done here.

Despite the fact that there already are formal ontologies dealing with mental disorders (SNOMED is such an example), we decided to develop a new one trying to overcome some of their potential limitations. In particular, with respect to the existing taxonomies of mental disorders, that are typically part of larger representation systems and must thus meet their criteria and general principles, we are currently building a representation that aims to be independent to previous ones and closer to the DSM-5 nosology and rationales (having also well clear in mind what are the main limits and problems of the Statistical Manual). We believe it is a necessary stage in order to verify and evaluate the exact limits of a classical approach to the formal representation of the concepts of individual mental disorders, as we suspect that some problems encountered by the already available formal ontologies might be due to an oversimplification of the structure and rationales of DSM-5 descriptive nosology, which is syndrome based and criterial.

On this respect, we take seriously the DSM-5 definition of the concept of MENTAL DISORDER, according to which a mental disorder is primarily a syndrome, that is a set of symptoms and signs. This means, for example, that the classes of *Mental_Disorder* and *Symptom* must be linked through an appropriate property (it must be remembered that, in OWL terminology, *properties* correspond to roles, or two place relations). Making the relationships between mental disorders and pattern of symptoms explicit might also help to clear out some classification disputes about where to place some controversial mental disorders among DSM-5 chapters.

Moreover, even if the DSM-5 definition of the concept of MENTAL DISORDER requires a dysfunction being in place, there is also the widespread conviction that syndromes should not be reified. The possibility to discover that basic dimensions

Fig. 10.1 The process of building a DSM-5 compliant ontology

of functioning, and thus dysfunctioning, cut across traditional syndrome-based diagnostic categories is actually envisaged – as the NIMH Research Domain Criteria (RDoC) project seems to corroborate. This means, for example, that the class of *Mental_Disorder* must be conceived in non realist terms and the concept of MENTAL DISORDER clearly distinguished from the concept of DISEASE.

Broadly speaking, the rationale we have followed to build our DSM-5 compliant ontology can be summarized in 4 steps, as shown in Fig. 10.1 above:

1. Identification of main concepts;
2. Formalization of classes and properties;
3. Implementation;
4. Comparison between symptoms and evaluation (i.e. modeling decision about the taxonomical position and the related axioms that need to be added).

The goal of the first phase was identifying, organizing and structuring all the main concepts of the domain by using an abstract model, e.g. graphs or schemes. Initially, we focused on the chapter of Schizophrenia Spectrum only, and defined relevant concepts and properties through a glossary or dictionary written in natural language. Afterwards, with the second phase we used description logics to formalize all the concepts and properties previously identified and thus to obtain the adequate terminological domain knowledge. The third phase aimed at encoding and implementing a formal ontology using *Protégé*, a widespread ontology editor developed at the Stanford University (http://protege.stanford.edu/). In the fourth and last phase we compared various symptoms among ontologies and different disorders. Moreover, the process of evaluation can be also driven in parallel with the previous three steps.

Fig. 10.2 Top level of the schizophrenia spectrum ontology

The Schizophrenia_Spectrum ontology that we have developed is currently composed by 58 classes, 5 properties and 191 axioms. As already mentioned, the ontology has been developed by adopting the OWL-DL (Ontology Web Language-Description Logic) dialect.

The three main classes are *Mental_Disorder*, *Patient* and *Symptom*. The top level of the ontology, which focuses on the various classes of the Schizophrenia Spectrum category and the associated symptoms, is shown in Fig. 10.2.

The top-level classes chosen by the adopted modeling not only allow that each mental disorder might be identified through quite different set of symptoms (as it is clearly demanded by the DSM-5 diagnostic criteria), but also address comorbidity (a phenomenon which is still common in DSM-5 meaning that each patient showing a certain set of symptoms might be diagnosed with more than one mental disorder).

The class of *Patient* allows to model many different patient instances, which is useful in order to include personal information regarding individuals (such as age, sex, gender, ethnicity, etcetera).

Finally, the class *Symptom* currently contains the following main subclasses, which have been built in accordance with the DSM-5 criteria: *Delusions*, *Disorga-*

nized Thinking, *Grossly Disorganized Abnormal Motor Behavior*, *Hallucinations*, *Negative Symptoms*.

The class of *Symptom* and its subclasses are disjointed from the *Schizophrenia Spectrum other Psychotic Disorder* and its subclasses as this guarantees the separation from symptoms that involves other mental disease.

The current version of the ontology (which is still subject to revisions and extensions) is available in a navigable format at http://www.di.unito.it/~lieto/Schizophrenia_Spectrum.html and downloadable at http://www.di.unito.it/~lieto/Schizophrenia_Spectrum.owl. Even if we developed a formal ontology which is more DSM-5 compliant than others (for instance, Ceusters and Smith 2010), as we predicted it is still unable to handle the representation and reasoning of common-sense cues.

10.5 A Proposal: A Hybrid Architecture

In this perspective – given the fact that the concepts of MENTAL DISORDER and individual mental disorders clearly exhibit typicality effects that cannot be handled with traditional, purely compositional, representational systems – we propose to integrate typicality effects in computational representations of concepts. More precisely, we focus on prototypical and exemplar based approaches, and propose to combine them in a hybrid model. (For the time being, we do not take into consideration here theory-theory approaches, since them are in some sense more vaguely defined if compared to the other two positions.)

Following the approach proposed in Frixione and Lieto (2013, 2014b) and preliminary tested in Lieto et al. (2015), we propose a hybrid architecture (see Fig. 10.3) combining a "classical" component (in which concepts are represented, as far as it is possible, in terms of necessary and/or sufficient conditions) with a "typicality-oriented" component, allowing both prototype and exemplar-based representations.

The "classical" component is demanded to some standard ontological formalism, such as DLs; the "typicality-oriented" component to a conceptual space, where conceptual spaces are a geometric framework for knowledge representation proposed by Peter Gärdenfors (2014).

In a conceptual space concepts are described in terms of a number of quality dimensions. In some cases, such dimensions are directly related to perception; examples could be temperature, weight, brightness, pitch. In other cases, dimensions can be more abstract in nature. To each quality dimension is associated a geometrical (topological or metrical) structure.

The central idea behind this approach is that the representation of knowledge can take advantage from the geometrical structure of the spaces. Instances (or exemplars) are represented as points in a space, and their degree of similarity can be calculated in a natural way according to some suitable distance measure. Concepts correspond to regions, and regions with different geometrical properties

Fig. 10.3 The conceptual architecture we propose

correspond to different kinds of concepts. Prototypes and typicality effects have a natural geometrical interpretation: a prototype corresponds to the geometrical centre of the region representing a concept (provided that the concept corresponds to a convex region). Thus, given a concept, a degree of centrality can be associated to each point that falls within the corresponding region. This degree of centrality can be interpreted as a measure of its typicality. Conversely, given a set of n prototypes represented as points in a conceptual space, a tessellation of the space in n convex regions can be determined in the terms of the so-called Voronoi diagrams. An example is shown in Fig. 10.4, where the center of each region corresponds to the prototype of a given concept, and where different exemplars can be represented as points in a conceptual region. The similarity between exemplars, or between prototypes and exemplars is obtained by calculating the metric distances in the underlying space.

In sum, the appeal of conceptual spaces consists in the fact that they provides a natural way of representing typicality effects, and that their geometrical structure provides a natural way of calculating the semantic relations between concepts, prototypes and exemplars in terms of metrical distance. In general, conceptual spaces seem to provide a better framework for modeling typicality effects in artificial system if compared to both standard symbolic systems and connectionist architectures – on this aspect see Lieto et al. (2017).

Considering the concepts of MENTAL DISORDER and individual mental disorders an hybrid architecture as the one described above would result particularly useful. On the one hand, the "classical" component – demanded to the ontology we developed and described in the above section – would allow us to make important inferences and comparisons between individual mental disorders. Moreover, it

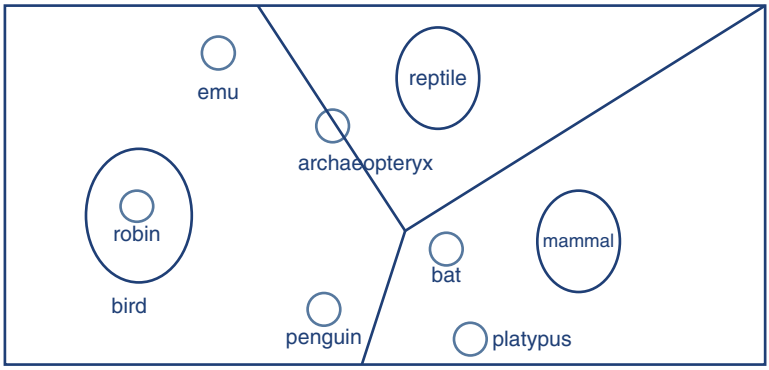

Fig. 10.4 An example of the Voronoi tessellation of a conceptual spaces. (From Gardenfors and Williams 2001)

would be needed to clearly represent some individual mental disorders that seem to be characterized by necessary (and sufficient) conditions (their medical or substance aetiology or some specific symptoms). For example, Bulimia nervosa seems to be characterized by necessary and sufficient conditions, as shown by its diagnostic criteria below:

A. Recurrent episodes of binge eating. An episode of binge eating is characterized by both of the following:
   1. Eating, in a discrete period of time (e.g., within any 2-hour period), an amount of food that is definitely larger than what most individuals would eat in a similar period of time under similar circumstances.
   2. A sense of lack of control over eating during the episode (e.g., a feeling that one cannot stop eating or control what or how much one is eating).

B. Recurrent inappropriate compensatory behaviors in order to prevent weight gain, such as self-induced vomiting; misuse of laxatives, diuretics, or other medications; fasting; or excessive exercise.
C. The binge eating and inappropriate compensatory behaviors both occur, on average, at least once a week for 3 months.
D. Self-evaluation is unduly influenced by body shape and weight.
E. The disturbance does not occur exclusively during episodes of anorexia nervosa. (American Psychiatric Association 2013, p. 345).

Alternatively, Major neurocognitive disorders are characterized by similar syndromes and differentiated by their underline pathological cause (e.g., Alzheimer's disease, Frontotemporal lobar degeneration, Lewy body disease, Vascular disease, Traumatic brain injury, HIV infection, Prion disease, Parkinson's disease, Huntington's disease), which is thus a necessary condition for their diagnosis.

Also the general concept of MENTAL DISORDER, at least its theoretical and "conceptually clean" version – as the one stated in the Introduction of DSM-5 – may allow for necessary (and sufficient) criteria, such as the dysfunction requirement:

> A mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that *reflects a dysfunction* in the psychological, biological, or developmental processes underlying mental functioning. Mental disorders are usually associated with significant distress or disability in social, occupational, or other important activities (American Psychiatric Association 2013, p. 20).

On the other hand, the "typicality-oriented" component would be necessary to deal with typicality effects and handle all those individual mental disorders, such as those listed in the chapter "Schizophrenia Spectrum and Other Psychotic Disorders", that are not characterized by necessary and sufficient conditions. Moreover, as far as the general concept of MENTAL DISORDER is concerned, the "typically-oriented" component would be useful to represent its common sense or practical version, which is much needed to guide us in distinguishing between health and pathological conditions in most of ordinary situations (Amoretti et al. 2017).

On this respect, we shall try to develop a conceptual space with a number of quality dimensions able to identify the prototype of mental disorder as well as the relevant exemplars. Some candidates for the qualitative dimensions might be related to the duration of symptoms, their clinical significance, their functional dimensions, and so on. Such as geometric framework would constitute the "non-classical" component of our architecture. Then, we shall also try to represent the various concepts of individual mental disorders within such as conceptual space and evaluate their positions, as well as their degree of typicality.

10.6 Concluding Remarks

To sum up, we exposed the problems raised by conceptual "typicality" to the classical theory of concepts, focusing on the general concept of MENTAL DISORDER and the various concepts of individual mental disorders as described by DSM-5. Then, we summarized one important issue faced by description logics in representing medical knowledge: as they are associated to a model theoretic, Tarskian style semantics, they prove to be unable to represent concepts in prototypical terms. To reinforce this conclusion, we build an ontology specifically suited to represent the general concept of MENTAL DISORDER and (most of) the various concepts of individual mental disorders. Despite being more DSM-5 compliant than other ontologies, our formalism was still unable to handle typicality effects. We thus proposes a hybrid approach combining a "classical" component (in which concepts are represented in terms of necessary and/or sufficient conditions) with a "typicality-oriented" component, allowing both prototype and exemplar-based representations.

In order to develop such a hybrid architecture, the next step would be defining a suitable number of quality dimensions able to identify the prototype of the general concept of MENTAL DISORDER and its relevant exemplars (such as SCHIZOPHRENIA, BORDERLINE PERSONALITY DISORDER, or MAJOR DEPRESSION). One possible application of this integration would be the real-

ization of an artificial system that, given a set of typical traits characterizing the different symptoms, would be able to provide the identification of the corresponding mental disorder.

References

American Psychiatric Association. 1952. *Diagnostic and Statistical Manual. Mental Disorders*. Washington, DC: American Psychiatric Association.

———. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: DSM-5*. Washington, DC: American Psychiatric Publishing.

Amoretti, M.C. 2015. *Filosofia e medicina. Pensare la salute e la malattia*. Roma: Carocci.

Amoretti, M.C., M. Frixione, and A. Lieto. 2017. The benefits of prototypes: The case of medical concepts. *Reti, Saperi e Linguaggi. Italian Journal of Cognitive Sciences* 4 (1): 97–114.

Boorse, C. 1976. What a theory of mental health should be. *Journal for the Theory of Social Behaviour* 6: 61–84.

Brachman, R., and J.G. Schmolze. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science* 9: 171–216.

Ceusters, W., and B. Smith. 2010. Foundations for a realist ontology of mental disease. *Journal of Biomedical Semantics* 1 (1): 10.

Cooper, R. 2007. *Psychiatry and Philosophy of Science*. Montreal: McGill-Queen's University Press.

Fodor, J. 1981. The present status of the innateness controversy. In *Representations*, ed. J. Fodor. Cambridge, MA: MIT Press.

Frixione, M., and A. Lieto. 2012. Representing concepts in formal ontologies: Compositionality vs typicality effects. *Logic and Logical Philosophy* 21: 391–414.

———. 2013. Dealing with concepts: From cognitive psychology to knowledge representation. *Frontiers in Psychological and Behavioural Science* 2 (3): 96–106.

———. 2014a. Concepts, perception and the dual process theories of mind. *Baltic International Yearbook of Cognition, Logic and Communication* 9: 1–20.

———. 2014b. Towards an extended model of conceptual representations in formal ontologies: A typicality-based proposal. *Journal of Universal Computer Science* 20 (3): 257–276.

Galatzer-Levy, I.R., and R.A. Bryant. 2013. 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science* 8 (6): 651–662.

Gärdenfors, P. 2014. *The Geometry of Meaning. Semantics Based on Conceptual Spaces*. Boston: MIT Press.

Gärdenfors, P., and M.-A. Williams. 2001. Reasoning about categories in conceptual spaces. *Proceedings IJCAI 2001*, 385–392.

Guarino, N. 1998. Formal ontology in information systems. *Proceedings of the First International Conference on Formal Ontologies in Information Systems (FOIS'98), June 6–8, Trento, Italy*, 46. Amsterdam: IOS Press.

Horrocks, I., P.F. Patel-Schneider, and F. Van Harmelen. 2003. From shiq and rdf to owl: The making of a web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web* 1 (1): 7–26.

Lieto, A., A. Minieri, A. Piana, and D. Radicioni. 2015. A knowledge-based system for prototypical reasoning. *Connection Science* 27 (2): 137–152.

Lieto, A., A. Chella, and M. Frixione. 2017. Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically inspired cognitive architecture*, 19, 1–9.

Lilienfeld, S.O., and L. Marino. 1995. Mental disorder as a Roschian concept: A critique of Wakefield's 'harmful dysfunction' analysis. *Journal of Abnormal Psychology* 104 (3): 411–420.

———. 1999. Essentialism revisited: Evolutionary theory and the concept of mental disorder. *Journal of Abnormal Psychology* 108 (3): 400–411.

Machery, E. 2009. *Doing without concepts*. Oxford: Oxford University Press.

McNally, R. 2011. *What is Mental Illness?* Cambridge, MA: Belknap Press of Harvard University Press.

Murphy, G. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Pickering, N. 2013. Extending disorder: Essentialism, family resemblance and secondary sense. *Medicine, Health Care and Philosophy* 16 (2): 185–195.

———. 2016. Disease, variety, disagreement, and typicality. Advantage Roschian concepts? *Philosophy, Psychiatry and Psychology* 23 (1): 17–31.

Rosch, E. 1975. Cognitive representation of semantic categories. *Journal of Experimental Psychology* 104: 573–605.

Rosch, E., and C.B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7 (4): 573–605.

Sadegh-Zadeh, K. 2000. Fuzzy health, illness, and disease. *Journal of Medicine and Philosophy* 25 (5): 605–638.

———. 2008. The prototype resemblance theory of disease. *Journal of Medicine and Philosophy* 33 (2): 106–139.

———. 2011. *Handbook of analytic philosophy of medicine*. New York: Springer.

Seising, R., and M. Tabacchi. 2013. *Fuzziness and medicine*. Dordrecht: Springer.

Wakefield, J.C. 1992. The concept of mental disorder. On the boundary between biological facts and social values. *American Psychologist* 47: 373–388.

———. 1999. Evolutionary versus prototype analyses of the concept of disorder. *Journal of Abnormal Psychology* 108 (3): 374–399.

Wittgenstein, L. 1953. *Philosophische Untersuchungen/Philosophical investigations*. Oxford: Blackwell.

Chapter 11
Large-Scale Simulations of the Brain: Is There a "Right" Level of Detail?

Edoardo Datteri

Abstract A number of research projects have recently taken up the challenge of formulating large-scale models of brain mechanisms at unprecedented levels of detail. These research enterprises have raised lively debates in the press and in the scientific and philosophical literature, some of them revolving around the question whether the incorporation of so many details in a theoretical model and in a computer simulations of it is really needed for the model to be explanatory. Is there a "right" level of detail? In this article I analyse the claim, made by two leading neuroscientists, according to which the content of the why-question addressed and the amount of computational resources available constrains the choice of the most appropriate level of detail in brain modelling. Based on the recent philosophical literature on (neuro)scientific explanation, I distinguish between two kinds of details, called here mechanistic decomposition and property details, and argue that the nature of the why-question provides only partial constraints to the choice of the most appropriate level of detail under the two interpretations of the term considered here.

Keywords Neuroscience · Computer simulation · Neural modeling · Brain project · Mechanistic decomposition · Levels of analysis

11.1 Introduction

In a *Scientific American* article entitled "An imitation of life", published in 1950 (Grey Walter 1950), pioneer of electroencephalography William Grey Walter described two small mobile robots, built by himself and named Elmer and Elsie, which were able to wander without hitting obstacles in the dark and to steer

E. Datteri (✉)
Department of Human Sciences for Education, RobotiCSS Lab – Laboratory of Robotics for the Cognitive and Social Sciences, University of Milano-Bicocca, Milan, Italy
e-mail: edoardo.datteri@unimib.it

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_11

towards light sources. In the article, Grey Walter stressed the fact that Elmer's and Elsie's reactive, apparently purposeful, and relatively unpredictable behaviour was produced by an extremely simple mechanism, and suggested, in quite metaphorical terms, that "only a few richly interconnected elements can provide practically infinite modes of existence" (p. 44). The fact that very simple mechanisms[1] can produce practically unpredictable behaviours when situated in non-structured, realistic environments, has been more recently emphasized by neuroscientist Valentino Braitenberg (Braitenberg 1986) and by roboticists Rodney Brooks (Brooks 1991) and Rolf Pfeifer (Pfeifer and Bongard 2006; Pfeifer and Scheier 1999), among others.

To be sure, immediately after the passage quoted above, Grey Walter surmises that mechanisms richer in components than those implemented in Elmer and Elsie may be required to explain phenomena such as "our subjective conviction of freedom of will and our objective awareness of personality" (Grey Walter 1950, p. 44). Indeed, what we can learn from Grey Walter's suggestions and tortoises is not that the "keep it simple" strategy will always succeed in the explanation of mental and behavioural phenomena. More properly, from his works we can derive a regulative principle that can be profitably followed in many processes of theorization over human and animal behaviour: one should consider, first, the possibility that the behaviour under investigation results from the interaction of simple mechanisms with real-world environments (see also Simon 1996). One should start from simple mechanisms, and add complexity only if they do not provide satisfactory predictive and explanatory grounds. Grey Walter's tortoises, Braitenberg's vehicles, and the robots built by Brooks and Pfeifer have been often cited in the (philosophy of) cognitive science and neuroscience literature to support this regulative principle, which is one of the central pillars of the so-called "situated cognition" movement (Cordeschi 2002; Tamburrini and Datteri 2005).

A number of recently funded research projects are apparently based on a very different regulative principle, which Komer and Eliasmith (2016) express in the following terms.

> Addressing cognitive behaviors in a neural model typically requires a large-scale model: one simulating tens of thousands to several million neurons. This is due to the correlation between the complexity in a task and its likelihood of being deemed 'cognitive.' Complexity suggests that large numbers of neural resources, and often many different brain areas, are required to address the challenge embodied by the task. (Komer and Eliasmith 2016, p. 14)

[1] Mechanisms are often said to be more or less "simple" (or "complex") in the cognitive science and neuroscience literature depending on the number of their components and on the number and nature of the connections among the components. The adjectives "simple" and "complex" are also often used to qualify behaviours, a complex behaviour being one which is relatively difficult to predict without computational instruments. Providing a precise definition of these terms is out of the scope of this article: here they will be used in the common-sense interpretation sketched here, just for the purpose of introducing the subject of the paper. In the following pages they will be abandoned, and mechanisms will be said to be more or less detailed according to a more precisely defined notion of "level of detail".

Under a reasonable interpretation, this passage – and other statements that can be found in the recent literature on large-scale modelling projects – implies that, to arrive at a "good" explanatory and predictive theoretical model of various human-level aspects of behaviour, one should start from models comprising an extremely large number of components. This regulative principle appears to be quite different from the regulative principle discussed before, and the research projects flowing from it may well be taken as exemplifying an emerging, important and reasonably well circumscribed approach in the theoretical modelling of animal and human behaviour. These projects, aiming to formulate large-scale models of brain mechanisms at unprecedented levels of detail, notably include the Blue Brain Project (Markram 2006), the Human Brain Project (Markram et al. 2011), the Cognitive Computing via Synaptronics and Supercomputing project (Ananthanarayanan et al. 2009), and the Cognitive Computation project (Eliasmith et al. 2012; Komer and Eliasmith 2016). They have been listed among the most important worldwide initiatives to advance brain research in (Grillner et al. 2016).

These research enterprises have raised lively debates in the press and in the scientific and philosophical literature (Datteri and Laudisa 2016; Eliasmith and Trujillo 2014; Miłkowski 2015). Many of these debates revolve around the question whether the incorporation of so many details in a theoretical model and in a computer simulations of it – which requires great modelling efforts and large amounts of funds – is really needed for the model itself to be explanatory. Notably, in a recent publication, two leading researchers (Eliasmith and Trujillo 2014) have raised the question whether there is a "right" level of detail to be pursued in brain modelling, and their answer is as follows.

> Is there a right 'level of detail'? We believe that this is simply an ill-posed question. As has long been accepted by those constructing large-scale climate models, the appropriate scale is determined by balancing two things … first, the questions that need to be answered and second, the available computational resources. If we are asking questions about how changing the morphology of neurons relates to changes in its activity (perhaps to understand the effects of neurofibrillary tangles in Alzheimer's disease), our model likely needs to include neuron morphology. However, if we are asking questions about how neuronal death in hippocampus results in memory loss, perhaps our model can simplify away detailed morphology. The benefit of such simplifications is that we can simulate more neurons using the same computational resources. (Eliasmith and Trujillo 2014, p. 1)

While believing that large-scale models are required to understand human behaviour in all its complexity, Eliasmith and Trujillo suggest that one should not abuse of the possibility – opened by the development of novel supercomputers, like the "Blue Gene" used in the Blue Brain Project (Markram 2006) – of formulating hyper-detailed models of the brain. Indeed, according to the two authors, there is not an absolutely "right" level of detail. The appropriate level of detail has instead to be chosen on a case-by-case basis, implying that hyper-detailed modelling is not always the best choice. The decision has to be made considering two factors, namely, the content of the why-question addressed and the computational resources available.

Their epistemological proposal is of significant interest. It may provide a guideline for choosing the most appropriate level of detail in the formulation of predictive and explanatory models of the brain. It may also provide criteria to assess the epistemic value of existing modelling and simulative research enterprises – more precisely, to assess whether the level of detail pursued in particular research projects, and the amount of funds and resources allocated for securing that level of detail, are appropriate. The goal of this paper is to reflect on whether the epistemological proposal made by Eliasmith and Trujillo is satisfactory – i.e., on whether the two factors proposed by them can really guide one in the choice of the appropriate level of detail in brain modelling.

I will argue that the choice of the appropriate level of detail cannot be based on one of the two factors proposed by the authors, that is to say, on the amount of computational resources available. I will also argue that whether the content of the why-question addressed determines the appropriate level of detail of a theoretical model is a question that depends (1) on what exactly we mean with "level of detail", and (2) on the type of the why-question itself. In Sect. 11.2, based on some epistemological statements made by Blue Brain Project leader Henry Markram (2006), I argue that that contemporary large-scale simulation projects aim to build hyper-detailed models of the brain under at least two interpretations of the notion of "detail": I will distinguish between mechanistic decomposition and property details. In Sect. 11.3, borrowing from (Craver 2002), I will distinguish between same-level and intra-level why-questions, and use this distinction to reflect upon the conditions under which the content of same-level and intra-level why-questions may constrain the choice of an appropriate level of detail:

- I will argue that the content of intra-level questions places stronger constraints than the content of same-level questions on the choice of the "right" level of mechanistic decomposition details;
- I will also argue that the content of same-level and intra-level questions places only partial constraints, in a sense to be discussed, on the choice of the most appropriate level of property details.

In all these cases, however, auxiliary epistemic principles are needed to choose how many mechanistic decomposition and property details the model should have I believe that the tentative observations sketched here may make some contribution towards an understanding of the regulative principles and the epistemic criteria guiding the formulation of predictive and explanatory mechanistic models of the brain.

11.2 Mechanistic Decomposition and Property Details in Large-Scale Neural Modelling

11.2.1 The Essential Building Blocks to Reconstruct Neural Circuitry

One of the most influential research projects which have recently aimed to formulate hyper-detailed mechanistic models of the brain, and to build hyper-accurate simulations of them, is the Blue Brain project led by neuroscientist Henry Markram. In his (Markram 2006, p. 155), he describes what he takes to be the "minimal essential building blocks required to reconstruct a neural microcircuit" (p. 155) in the following terms.

> To model neurons, the three-dimensional morphology, ion channel composition, and distributions and electrical properties of the different types of neuron are required, as well as the total numbers of neurons in the microcircuit and the relative proportions of the different types of neuron. To model synaptic connections, the physiological and pharmacological properties of the different types of synapse that connect any two types of neuron are required, in addition to statistics on which part of the axonal arborization is used (presynaptic innervation pattern) to contact which regions of the target neuron (postsynaptic innervation pattern), how many synapses are involved in forming connections, and the connectivity statistics between any two types of neuron (p. 155).

I take these claims to express the belief that the incorporation of high levels of details is to be praised in the mechanistic modelling of the brain – a belief that, as pointed out earlier, is at the basis of the approach adopted in the Blue Brain Project and in other projects. However, one should be careful to note that here Markram conflates various dimensions along which a neuroscientific mechanistic model can be said to be more or less detailed. In the following sub-sections I will try to distinguish between them.

11.2.2 Level of Mechanistic Decomposition

"Microcircuits are composed of neurons and synaptic connections" (p. 155). According to Markram (2006) (see also Fig. 2 of the article), in order to accurately model neurons one has to provide information on their electrophysiological behavior, on the composition and distribution of ion channels, and on the number of synaptic boutons (i.e., the specialized areas of the presynaptic terminal which contain the machinery required to release neurotransmitters into the vicinity of the post-synaptic neuron). What kind of details is Markram talking about in these claims?

As pointed out in the recent literature on mechanistic explanation, mechanistic models describe the regular behaviour of system components by means of generalizations (Craver 2007; Glennan 2002; Woodward 2002). This is true of

Fig. 11.1 The mechanistic decomposition hierarchy

neuroscientific mechanistic models too. Some neuroscientific models characterize the behaviour of selected neural areas, treated as closed-box components of a larger mechanism, by reference to the regular relationship between the average firing rate of input and output neural populations. As often pointed out in the literature on functional and mechanistic explanation, mechanistic analysis can be iterated on each component of a previously formulated mechanism (Piccinini and Craver 2011; Levy and Bechtel 2013; Cummins 1975; Glennan 2002; Rosenblueth and Wiener 1945). For instance, one can analyse the mechanism governing the input-output behaviour of neural areas which were previously treated as closed boxes and specify, say, the electrophysiological behaviour of some neurons included in it in terms of regular relationships between dendritic and axonal membrane voltages. By "opening the box" – that is to say, by iterating mechanistic analysis on the components of a larger mechanism – one goes downwards along what may be called a *mechanistic decomposition hierarchy* (see Fig. 11.1). In a sense, the mechanistic model positioned at level n-1 will be more detailed than the mechanism positioned at level n. The details added in this way will be called *mechanistic decomposition details* from now on.

Under a reasonable interpretation, the details Markram is talking about in the claims mentioned at the beginning of this sub-section are mechanistic decomposition details. By modelling the electrophysiological profile of the neurons included in a neural microcircuit, one describes the generalizations governing the components of that circuit. This amounts to carrying out mechanistic analysis of the neural microcircuit. Ion channels and synaptic boutons are parts of neurons: by modelling

their behaviour – taking into account ionic permeability, probability of neurotransmitter release, depression and facilitation time constants for each component (see Fig. 2 of Markram 2006, p. 155) – one carries out mechanistic analysis of the components of the neural microcircuit, thus taking a further step downwards along the mechanistic decomposition hierarchy. In this way, one progressively adds mechanistic decomposition details to the initial model. These are good reasons to claim that, in Markram's approach, explanatorily adequate models of neural microcircuits should be rich in mechanistic decomposition details.

11.2.3 Abstraction in the Characterization of Each Component

According to Markram (2006), "to model neurons, the three-dimensional morphology ... of the different types of neuron are required ... To model synaptic connections, the physiological and pharmacological properties of the different types of synapse that connect any two types of neuron are required" (p. 155). Synaptic behaviour must also be characterized, in Markram's approach, taking into account the pre- and post-synaptic innervation patterns. By adding previously omitted information on the morphology of neurons, and on the physiological and pharmacological properties of synapses, one obviously add details to a neural model. However, these details are different in nature from the mechanistic decomposition details discussed in the previous section.

To understand, suppose that the behaviour of a purely notional component c_1 of mechanism M is modelled in terms of a relationship between the values, say, of electrical properties p_1 and q_1, an example being an oversimplified model neuron characterized only in terms of the relationship between the average value of dendritic membrane potential (p_1) and the potential at a specific point in the axon (q_1). A first way to add details to this simple model is to increase the number of the electrical properties taken into account. For instance, one may characterize the behaviour of the neuron in terms of a relationship between the membrane potential at *many* points in the dendritic structure and the axonal membrane potential, thus in terms of a many-to-one relationship between electrical properties $p_1, \ldots, p_n$ and electrical property q_1. Increasing the number of "input" electrical variables would be required if one intended to describe the pattern of connections from presynaptic neurons to the target neuron (see Markram's claims summarized at the beginning of this sub-section). Such a model would be clearly more detailed than the oversimplified model I have mentioned before.

A second case, which is slightly but interestingly different from the previous one, is when one adds properties expressed in different theoretical vocabularies. Suppose, for example, that one adds *morphological* information on a neuron which was had been previously characterized only in terms of its *electrophysiological* behaviour (see Markram's claims above). In a sense, the resulting model is more detailed than the previous one.

As a third case, suppose that one adds information on the boundary conditions under which the behaviour of c_1 is expected to behave according to a given regularity. In brain models, boundary conditions are typically expressed in terms of other (environmental or physiological) properties whose value must be higher or lower than a particular threshold, or within a given range. A case in point is when one adds *pharmacological* details to the model (see Markram's claims above), e.g., by stating that a neuron will fire only if the extracellular concentration of a particular molecule is below some given threshold.

Note that the details discussed in these three examples are not of the mechanistic decomposition variety. By adding them, one enriches the way component c_1 is described without carrying out a mechanistic analysis of it. I will call *property* details the information added in these examples, to distinguish them from the mechanistic decomposition details discussed in Sect. 11.2.2. Accordingly, I take the claims mentioned at the beginning of this sub-section as implying that, according to Henry Markram, "good" models of neural microcircuits should be rich in property details.

Markram's comments on the Cognitive Computing via Synaptronics and Supercomputing project led by IBM researcher Dharmendra Modha (Ananthanarayanan et al. 2009) plausibly rest on the claim that predictively and explanatorily adequate models of the brain must be rich in property details. In the framework of that project a massively parallel cortical simulator, called C2, has been built "with 1.617×109 neurons and 0.887×1013 synapses, roughly 643 times slower than real-time per Hertz of average neuronal firing rate. The model used biologically-measured gray matter thalamocortical connectivity from cat visual cortex" (Ananthanarayanan et al. 2009, p. 1). In a letter sent to the IBM Chief Technical Officer in 2009, Markram claimed that it was "shameful and unethical" to call C2 a simulation of the cat's brain. His point was that the neuron models simulated by Modha "are point neurons [with] no branches; [...] the simplest possible equation you can imagine to simulate a neuron, totally trivial synapses, [...]. All these kinds of simulations are trivial and have been around for decades – simply called artificial neural network (ANN) simulations. [...] It is really no big deal to simulate a billion points interacting if you have a big enough computer".[2] Markram's reaction concerns the level of property detail of the base components of the mechanism. One thing is to simulate a point neuron with, say, a linear input-output function and few dendrites; another thing is to simulate neurons with a higher degree of dendritic and axonal arborisation, and with nonlinear input-output characteristics. What changes from the former to the latter model is the number and type of property details.

[2] http://spectrum.ieee.org/tech-talk/semiconductors/devices/blue-brain-project-leader-angry-about-cat-brain (last visited on September 14, 2016).

11.3 On the Choice of the "Right" Level of Detail

11.3.1 Same-Level and Inter-Level Questions

In the previous section I have separated two dimensions, apparently conflated by Markram (2006), along which a theoretical model of the brain can be said to be more or less detailed. This distinction can be useful to address the epistemological question addressed by Eliasmith and Trujillo (2014) in the passage quoted above: how to choose the "right" level of detail in the formulation of a theoretical model of the brain? Their answer is that the most appropriate level can be determined based on the nature of the why-question and the available computational resources. Here I will provide some insight to reflect on whether, and under what auxiliary epistemological assumptions, this solution is viable under the two interpretations of the notion of "detail" sketched before. Are the factors proposed by Eliasmith and Trujillo really helpful to choose the most appropriate level of detail of mechanistic decomposition (Sect. 11.2.2) and property (Sect. 11.2.3) details?

Arguably, one of them – the amount of computational resource available – is to be dropped, at least under a plausible interpretation of the term "appropriate". I assume that the epistemological question under scrutiny concerns the level of detail that a theoretical model should exhibit for it to be explanatory. The amount of available computational resources is surely to be taken into account to decide if the model can be accurately simulated or not in a computer. But the question whether a model can be accurately simulated or not is totally independent of the question whether that model is explanatory or not: the amount of available computational resources has no bearing whatsoever on the latter question. For this reason, under this interpretation of the term "appropriate", in what follows I will focus only on the relationship between the content of the addressed why-question and the choice of the appropriate level of detail of a brain model.

To prepare the ground for the following discussion, let me distinguish between (what I call here) same-level and inter-level why-questions, where the term "level" is to be understood as referring to the level of mechanistic decomposition (Sect, 11.2.2; see also Craver 2002 on inter-level questions).[3] Consider a purely notional neuroscientific system S, and the following sequence of theoretical models of it differing from one another in the level of mechanistic decomposition.

– Level 0: system S is treated as a closed box. The model describes its behaviour only in terms of a relationship between sensory stimuli and motor reactions, without specifying the mechanism connecting S′ sensors to motor organs.

[3] Several accounts of the formal semantics of why-questions can be found in the philosophical literature, most notably in (Bromberger 1966; Van Fraassen 1980; Hintikka and Halonen 1995). In what follows, I assume that the distinction between same-level and inter-level questions made here is compatible with all these accounts. Examples of why-questions are provided below in the text.

– Level -1: the mechanism connecting S′ sensors and motor organs is described, the base components of it being neural areas. Each neural area is treated as a closed box, and its behaviour is characterized in terms of a relationship between the firing rate of particular "input" and "output" neurons.
– Level -2: each neural area mentioned in the level -1 model is analysed in terms of a mechanism whose base components are neurons. Each neuron is treated as a closed box, and its behaviour is characterized electrophysiologically in terms of a number of relationships between dendritic and axonal membrane voltages.
– Level -3: each neuron mentioned in the level -2 model is analysed in terms of a mechanism whose base components include ion channels. The behaviour of each ion channel is characterized in terms of a relationship between chemical conditions (including the intra- and extracellular concentration of specific ions, the presence of ligands, and the membrane voltage) and the openness/closeness of the channel pore.

With reference to this hierarchy of models, consider the two following why-questions:

1. Why is the firing rate of the output neurons of area c_1 such and such whenever c_1's input neurons fire?
2. Why is S' motor behaviour such and such whenever c_1's ion channels bind to a particular molecule?

Each question asks for an explanation of a regularity. The first regularity links input and output firing activity of a particular neural area c_1. Neural area c_1 is a base component of the level -1 theoretical model in the above hierarchy. In a sense, therefore, question 1 asks for an explanation of a regularity pertaining to level -1 only: it is (what will be called here) a *same-level* why-question. Other examples of same-level questions are: why is S′ motor behaviour such and such whenever sensory stimulus *s* is delivered to the system (level 0)? Why does neuron n_1 of area c_1 produce a spike whenever the pattern of dendritic voltage is such and such (level -2)? Why does ion channel i_1 of neuron n_1 open whenever a particular molecule binds to the channel? Same-level questions need not be about exactly one component of the system. A level -2 question on the relationship between the dendritic voltage of neuron n_1 and the axonal voltage of neuron n_2 would count as a same-level question: in a sense, n_1 and n_2 are base components of the same theoretical model in the hierarchy – they are same-level components.

Regularity 2 links S′ motor behaviour with the activity of c_1's ion channels. S′ motor behaviour is the output of a base component of level 0 (which is system S itself, treated as a closed box). Ion channels are base components of the level -3 mechanism. In a sense, question 2 asks for an explanation of a regularity which spans different levels of the mechanistic decomposition hierarchy: it is an *inter-level* why-question. Other examples of inter-level question are: why is S′ motor behaviour (level 0) such and such whenever neuron n_1 produces a spike (level -2)? Why does neuron n_1 produce a spike (level -2) whenever n_1's ion channels bind to a specific

molecule (level -3)? Why is the output of neural area c_1 such and such (level -1) whenever the ion channels of c_1's neurons bind to a specific molecule (level -3)?

With respect to the last question, one should be careful to note that area c_1, which is a component of the level -1 model, is likely to be mentioned at level -3 too as the area whose neurons include the ion channels the question is about. More generally, one may reasonably claim that every theoretical model below level 0 will mention all the mechanistic decomposition details introduced at higher levels (as represented in Fig. 11.1). For this reason, one may be tempted to classify the question above as a same-level, level -3 question. However, note that area c_1 is a base component of the level -1 mechanism – it is treated as a closed box at that level – but not of the level -3 mechanism. It is in this sense that c_1 may be classified as a level -1 component, and that the why-question above may be classified as an inter-level, and not a same-level, question.[4]

11.3.2 Why-Questions and Levels of Analysis

The distinction between same-level and inter-level why-questions can be brought to bear on Eliasmith's and Trujillo's epistemological thesis on how to choose the most appropriate level of analysis in a brain model. In particular, in this section, I will offer some insight on whether the content of same-level and inter-level why-questions may constrain the choice of the appropriate level of mechanistic decomposition (Sect. 11.2.2) and property (Sect. 11.2.3) details.

I start from the relationship between the content of same-level questions and the level of mechanistic decomposition details. To explain mechanistically why some regularity obtains, one has to identify the mechanism producing it (Craver 2007). For instance, with reference to the hierarchy described above, to explain why S′ motor output is such and such when S′ sensory inputs are such and such (which is a same-level, level 0 question), one has to formulate a model which includes information taken from lower levels of mechanistic decompositions – which amounts to identifying the mechanism producing that regularity. Choosing the most appropriate level of mechanistic decomposition details consists in deciding "how deep" one should go downwards along the mechanistic decomposition hierarchy. Should one be satisfied with a mechanistic model mentioning S' neural areas and their interconnections (level -1), or should one iterate mechanistic analysis to lower levels, by modelling the behaviour of the neurons composing those areas (level -2) and the ion channels encompassed in those neurons (level -3)? It is not clear why the content of the why-question should tell one where to stop. One may choose to

[4]This argument is to be refined based on a formal account of the notion of "mechanistic decomposition level", which is out of the scope of this paper. Note that why-questions can be classified as same-level or inter-level only with respect to a particular mechanistic decomposition hierarchy. No why-question is "intrinsically" same-level or inter-level.

keep the model as simple as possible and explain S' level-0 behaviour on the basis of a level -1 mechanism. This choice, however, would be guided by a principle of epistemic parsimony which is not implied by the content of the why-question in any clear way.

Arguably, the content of inter-level questions places stronger constraints on the choice of the appropriate level of mechanistic decomposition details. Consider the second why-question discussed in the previous section: why is S′ motor behaviour such and such whenever the ion channels of neural area c_1 bind to a particular molecule? As pointed out before, this is an inter-level question asking for explanation of a regularity connecting levels 0 to −3 in the notional hierarchy sketched above. An appropriate answer to this question will have to describe a mechanism linking ion channel activity to S′ motor behaviour, that is to say, a multi-level mechanism reaching level -3 from level 0 in the hierarchy. These levels must be covered for the model to provide the appropriate explanatory resources. To be sure, a model reaching lower levels of mechanistic decomposition (for example, a model connecting level 0 to the level at which the chemical mechanisms governing ion channel activity are described) might offer the same theoretical resources. Similarly to the same-level question case, one may reasonably decide to stop at level -3 on the basis of a principle of epistemic parsimony not implied by the content of the why-question. However, in addition to that auxiliary principle, the content of that question – in virtue of its being inter-level – provides a reason to add mechanistic decomposition details at least down to level -3.

I turn now to the question whether the content of same-level and inter-level why-questions constrains the choice of the appropriate level of property details (Sect. 11.2.3). Consider two same-level why-questions, one concerning the one-to-one relationship between the *average* dendritic voltage and the firing rate of a neuron, the other concerning the many-to-one relationship between the voltage *at many different points* in the dendritic membrane and the firing rate of the same neuron. Arguably, the theoretical model formulated to address the second question must be more property-detailed than the model used to answer the first question: it has to describe a mechanism connecting neuronal firing rate to a richer set of input electrical properties. Note, however, that this relative richness in property-details concerns the way the "input" and "output" of the mechanism are characterized. The content of the why-question does not place constraints on the number and nature of the property details characterizing the *internal* components of the explanatory mechanism: other auxiliary assumptions, apparently not implied by the content of the why-question itself, will guide this choice.

Analogous observations can be made concerning the choice of the theoretical vocabulary in which the model is couched, and of the boundary conditions under which the target system is expected to behave according to the specified mechanism (Datteri and Laudisa 2014). Consider two same-level why-questions, one concerning the regularity between changes of two, say, morphological properties of a system, the other concerning the regularity between changes of two electrophysiological properties of the same system. The theoretical models providing answers for these two questions will have to explain regularities couched in different theoretical

vocabularies. In the first case, one will have to describe a mechanism whose inputs and outputs are morphological properties of the target system, while in the second case the inputs and outputs of the mechanism will be electrophysiological properties. The language in which the why-question is formulated constrains the choice of the language in which the inputs and outputs of the system are couched (Datteri and Laudisa 2014). There seems to be no principled reason, however, for sticking to the same theoretical vocabulary in the description of the internal components of the mechanism as well – one may choose to explain a morphological regularity based on the electrophysiological behaviour of internal components of the mechanism, if suitable bridges between the two theoretical vocabularies are available.

By comparison, consider two inter-level questions, one asking for explanation of a regularity between changes of one level 0 property and of one level -3 property, the other one asking for explanation of a regularity between changes of one level 0 property and a higher number of level -3 properties. The theoretical model used to address the second question will have to be more property-detailed in the description of its level -3 components than the theoretical model used to address the first question. Arguably, therefore, the content of inter-level questions places some constraint on the choice of the property details of the internal components of a theoretical model. However, in this case too, nothing prevents one to add further property details to the model: the content of the why-question provides no guideline to decide on this issue.

11.4 Summary and Concluding Remarks

Many contemporary research projects aim to formulate theoretical models of brain functions at unprecedented levels of detail, apparently under the assumption that the more detailed the model is, the higher its explanatory and predictive value will be. This assumption has been challenged by Eliasmith and Trujillo (2014), according to which there is an appropriate level of detail to be chosen on a case-by-case basis. In their opinion, the decision has to be made taking into account the content of the why-question addressed and the amount of computational resources available. This epistemological thesis has been discussed in this paper. First, I have argued that the amount of computational resources available may be brought to bear on the possibility of accurately simulating a model, but that it does not constrain the level of detail of the model itself. Second, based on a distinction between mechanistic decomposition and property details, and by same-level and inter-level why-questions, I have offered reasons to believe that the content of the why-question may in some cases constrain the choice of the appropriate level of detail of the model, but that auxiliary principles not implied by the content of the why-question itself are needed to decide.

The analysis provided here can be refined and extended in a number of ways. In particular, I have not focused on other dimensions along which a theoretical model can be said to be more or less detailed (Datteri and Laudisa 2016), one of them

being the size of the model. According to Markram, a "good" theoretical model of a neural microcircuit must describe "the total number of neurons in the microcircuit and the relative proportions of the different types of neurons" (Markram 2006, p. 155). Simulating neural networks comprising a huge number of neural unit is one of the main goals of contemporary large-scale simulation projects. For example, the Blue Brain Project aimed to build a simulation of a portion of the somatosensory cortex of the rat composed of about 10.000 neurons, while the Blue Gene – the supercomputer used in the Blue Brain experiments – was reported to be able to simulate a 100.000-neuron neural network. Eliasmith's SPAUN model (Eliasmith et al. 2012) comprises 2.5 million neurons. According to Markram, the development of computational techniques able to simulate the entire human brain with its 100 billion neurons would "provide a strong foundation for taking the next quantum step, to further increase the size of the modelled network to an unprecedented level" (Markram 2006, p. 154). In a sense, by increasing the number of the base components of a theoretical model, one makes it more detailed. And there are good reasons to believe that size details are different in kind from the details discussed in the previous sections. How to choose how large the model should be – in other words, how to choose the "right" level of size details – for it to be explanatory? I believe that a philosophical reflection on the criteria guiding (neuro)scientists in deciding how detailed their models should be, possibly along the preliminary dimensions sketched here, may contribute to achieving a deeper understanding of what makes a "good" neuroscientific explanation, which is one of the central goals of the philosophy of cognitive science and neuroscience.

References

Ananthanarayanan, R., S.K. Esser, H.D. Simon, and D.S. Modha. 2009. The cat is out of the bag: Cortical simulations with 109 neurons, 1013 synapses. In *High performance computing networking, storage and analysis, proceedings of the conference on, (c)*, 1–12. https://doi.org/10.1145/1654059.1654124.

Braitenberg, V. 1986. *Vehicles. Experiments in synthetic psychology*. Cambridge, MA: The MIT Press.

Bromberger, S. 1966. Why-questions. In *Mind and Cosmos: Essays in contemporary science and philosophy*, ed. R. Colodny, 68–111. Pittsburgh: University of Pittsburgh Press.

Brooks, R.A. 1991. New approaches to robotics. *Science* 253 (5025): 1227–1232. https://doi.org/10.1126/science.253.5025.1227.

Cordeschi, R. 2002. *The discovery of the artificial. Behavior, mind and machines before and beyond cybernetics*. Dordrecht: Springer. https://doi.org/10.1007/978-94-015-9870-5.

Craver, C.F. 2002. Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy of Science* 69: September), 83–September), 97.

Craver, C. 2007. *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Clarendon Press.

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

Datteri, E., and F. Laudisa. 2014. Box-and-arrow explanations need not be more abstract than neuroscientific mechanism descriptions. *Frontiers in Psychology* 5 (MAY): 1–10. https://doi.org/10.3389/fpsyg.2014.00464.

———. 2016. Large-scale simulations of brain mechanisms: Beyond the synthetic method. *Paradigmi* 3: 23–46. https://doi.org/10.3280/PARA2015-003003.

Eliasmith, C., and O. Trujillo. 2014. The use and abuse of large-scale brain models. *Current Opinion in Neurobiology* 25: 1–6. https://doi.org/10.1016/j.conb.2013.09.009.

Eliasmith, C., T.C. Stewart, X. Choo, T. Bekolay, T. DeWolf, C. Tang, and D. Rasmussen. 2012. A large-scale model of the functioning brain. *Science* 338 (6111): 1202–1205. https://doi.org/10.1126/science.1225266.

Glennan, S. 2002. Rethinking mechanistic explanation. *Philosophy of Science* 69 (S3): S342–S353. https://doi.org/10.1086/341857.

Grey Walter, W. 1950. An imitation of life. *Scientific American* 182 (5): 42–45.

Grillner, S., N. Ip, C. Koch, W. Koroshetz, H. Okano, M. Polachek, and M. Poo. 2016. Worldwide initiatives to advance brain research. *Nature* 19 (9): 1118–1122. https://doi.org/10.1038/nn.4371.

Hintikka, J., and I. Halonen. 1995. Semantics and pragmatics for why-questions. *Journal of Philosophy* 92 (12): 636–657.

Komer, B., and C. Eliasmith. 2016. A unified theoretical approach for biological cognition and learning. *Current Opinion in Behavioral Sciences* 11: 14–20. https://doi.org/10.1016/j.cobeha.2016.03.006.

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

Markram, H. 2006. The blue brain project. *Nature Reviews. Neuroscience* 7 (2): 153–160. https://doi.org/10.1038/nrn1848.

Markram, H., K. Meier, T. Lippert, S. Grillner, R. Frackowiak, S. Dehaene, A. Knoll, H. Sompolinsky, K. Verstreken, J. DeFelipe, S. Grant, J.P. Changeux, and A. Sariam. 2011. Introducing the human brain project. *Procedia Computer Science* 7: 39–42. https://doi.org/10.1016/j.procs.2011.12.015.

Miłkowski, M. 2015. Explanatory completeness and idealization in large brain simulations: A mechanistic perspective. *Synthese* 193: 1457–1478. https://doi.org/10.1007/s11229-015-0731-3.

Pfeifer, R., and J. Bongard. 2006. *How the body shapes the way we think. A new view of intelligence*. Cambridge, MA: The MIT Press.

Pfeifer, R., and C. Scheier. 1999. *Understanding Intelligence*. Cambridge, MA: The MIT Press.

Piccinini, G., and C. Craver. 2011. Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183: 283–311.

Rosenblueth, A., and N. Wiener. 1945. The role of models in science. *Philosophy of Science* 12 (4): 316–321. https://doi.org/10.1086/286874.

Simon, H.A. 1996. *The sciences of the artificial*. Cambridge, MA: The MIT Press.

Tamburrini, G., and E. Datteri. 2005. Machine experiments and theoretical modelling: From cybernetic methodology to neuro-robotics. *Minds and Machines* 15 (3–4): 335–358. https://doi.org/10.1007/s11023-005-2924-x.

Van Fraassen, B. 1980. *The scientific image*. Oxford: Clarendon Press.

Woodward, J. 2002. What is a mechanism? A counterfactual account. *Philosophy of Science* 69: S366–S377 JOUR.

Chapter 12
Virtual Information in the Light of Kant's Practical Reason

Matteo Vincenzo d'Alfonso

Abstract In (D'Agostino M, Floridi L, Synthese 167:271–315, 2009) the authors face the so-called "scandal of deduction" (Hintikka J, Logic, language games and information. Kantian themes in the philosophy of logic. Clarendon Press, Oxford, 1973). This lies in the fact that the Bar-Hillel and Carnap theory of semantic information implies that tautologies carry no information. Given that any mathematical demonstration and more in general every logical inference in a first-order language can be reduced to a tautology; this would imply, that demonstrations bring no fresh information at all.

Addressing this question (D'Agostino M, Floridi L, Synthese 167:271–315, 2009) offers both: (i) a logical model for a strictly analytical reasoning, where the conclusions depend just on the information explicitly present in the premises; and (ii) a proposal for the ranking of the informativeness of deductions according to their increasing recourse to so called "virtual information", namely information that is temporarily assumed but not contained in the premises.

In this paper I will focus on the status of virtual information in its connection with the Kantian philosophical spirit. Exploiting the standard Kantian difference between theoretical and practical reason, my aim is to show that the access to virtual information is due to what Kant calls *practical reason* rather then to the theoretical one, even though the effects of its deployment are purely theoretical, i.e. don't lead an agent to any moral action but just to acquiring new information.

Keywords Carnap-Bar-Hillel paradox · Semantic information · Scandal of deduction · Virtual information · Synthetic a-priori · Practical reason · Theoretical reason · Kant

M. V. d'Alfonso (✉)
Dipartimento di Studi Umanistici, Università di Ferrara, Ferrara, Italy
e-mail: dalfonso@unife.it

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_12

12.1 Introduction

Hintikka (1973) defines "scandal of deduction" the fact that, as inferences in a first order language can be reduced to a tautology, according to the standard semantic interpretation of information they don't bring to any real epistemic gain: the information conveyed by their conclusions are already implicitly contained in the premises. This can be seen as a complementary conclusion to the Carnap-Bar-Hillel paradox, asserting that "a self-contradictory sentence ... is regarded as carrying with it the most inclusive information" (Carnap and Bar-Hillel, 1953, 224). Since first-order logic is also the standard for the formalization of mathematics, this leads to the highly unpleasant and counterintuitive conclusion that proving theorems carries no information gain, and at best bears merely a psychological value: something implicitly contained in the premises or axioms becomes eventually explicit without bringing any substantial advantage.[1] This would be moreover the true meaning of the statement that logico-mathematical sentences are *analytic*. Hintikka intends to vindicate the layman's intuition that mathematic and logic can (also) be informative and he does it by introducing the distinction between two kind of information, depth and surface information:

> [D]epth information is the totality of information we can extract from a sentence by all the means that logic puts to our disposal. Surface information, on the contrary, is only that part of the total information, which the sentence gives us explicitly. It may be increased by logical operations. In fact, this notion of surface information seems to give us for the first time a clear-cut sense in which a valid logical or mathematical argument is not tautological but may increase the information we have. In first-order logic, valid logical inferences must be depth tautologies, but they are not all surface tautologies.[2]

According to this definition we could conclude that all the cases where the demonstration of a theorem makes use of a construction of any kind, e.g. when for the demonstration of the properties of a geometrical figure we need to draw lines that exceed the ones composing the given figure and that are eventually ignored in the conclusion, should be considered as augmenting our surface information hence synthetic a-priori.

D'Agostino and Floridi (2009) extend Hintikka's intuition to propositional logic that Hintikka still held for truly analytic. In order to do so they introduce the concept of "virtual information", namely information that is temporarily assumed

[1] Hempel 1945: "Since all mathematical proofs rest exclusively on logical deduction from certain postulates, it follows that a mathematical theorem, such as the Pythagorean theorem in Geometry, *asserts nothing that is objectively or theoretically new* as compared with the postulates from which it is derived, although *its content may well be psychologically new* in the sense that we were not aware of its being implicitly contained in the postulates" (my emphasis).mathematical theorem, such as the Pythagorean theorem in Geometry, *asserts nothing that is objectively or theoretically new* as compared with the postulates from which it is derived, although *its content may well be psychologically new* in the sense that we were not aware of its being implicitly contained in the postulates" (Hempel 1945, my emphasis).

[2] Hintikka 1973, p. 22.

for the sake of reasoning, but is not contained in the premises and hence eliminated in the conclusion. In doing so, they offer both: (i) a logical model for a strictly analytical reasoning, where the argument never makes use of information that is not even implicitly contained in the premises; and (ii) a proposal for the ranking of the informativeness of deductions according to their increasing recourse to so called "virtual information",[3] showing that it is possible to assess the amount of information carried by an inference according to the depth at which virtual information is used in the inference process.

An important philosophical aspect of this solution is that it mitigates the difference between analytic and synthetic reasoning. In doing so it disentangles this difference from the *a-priori* and *a-posteriori* distinction: only deductions with depth 0 (i. e. no recourse to virtual information) can be defined strictly analytic whereas all other inferences, although undoubtedly a-priori, show a certain degree of synthesis. For this reason D'Agostino (2013 and 2014) refers to his solution as to a Kantian one. Its Kantian flavour would namely depend on stating the existence of a *synthesis a-priori* already at the level of the propositional logic, as we can qualify theorems being synthetic a-priori if their demonstration makes use of virtual information.

In this paper I will focus on the status of virtual information in its connection with the Kantian philosophical spirit. Exploiting the standard Kantian difference between theoretical and practical reason, my aim is to show that the access to virtual information is due to what Kant calls *practical reason* rather then to the theoretical one, even though the effects of its deployment are purely theoretical, i.e. don't lead an agent to any moral action but just to acquiring new information.

My argument will then proceed as follows: Sect. 12.2 will offer an overview of D'Agostino and Floridi's proposal of synthesis a-priori and how such proposal relates to the notion of *virtual information* (12.2.1); hereafter I will shortly review D'Agostino's view of 'virtual information' using the example of the Sudoku game (12.2.2); Sect. 12.3.1 will be devoted to the distinction between theoretical and practical use of reason in Kant, showing how the core of the Practical Reason consists in our faculty of thinking in terms of "*ought to*" or "*should*" (in German: *Soll*) and act according to it. Eventually, Sect. 12.3.2 and 12.4 will go back to the use of virtual information and propose that the condition under which we access them is the use of the "*ought to*/*should*", hence the synthetic value of demonstrations relies on the use of practical reason in a theoretical environment.

12.2 What Is Virtual Information and Why It Matters in Propositional Logic

As an example of demonstration recurring to "virtual information" (D'Agostino, Floridi 2009) presents Euclid's proof of the Fourth Proposition in the First Book of

[3] See Sect. 12.2 below.

the *Elements*. Here it is shown, by superposition, that triangles with equal sides and equal angles are also equal to one another, as they can be placed on one another. Well, in case of triangles being a mirror image of each other, this method can be applied only assuming that two-dimensional figures, as triangles are, can rotate in a third dimension, which is not given in the premises: i.e. *flatlandians* (Abbott 1884) would never be able to demonstrate the theorem. Hence the information that two-dimensional figures are plunged in a three-dimensional space, though not explicitly stated neither in the premises, nor in the proof, has to be employed for supporting an unavoidable step of the demonstration, but it is also eventually ignored in the conclusions: i.e. *flatlandians* though unable to understand how the triangle could ever rotate, once this move has been "magically" done, are immediately aware of the conclusion. In the light of this proof-method D'Agostino and Floridi (2013) suggest to reassess the difference between analytic and synthetic in the sense Kant did:

> [T]he reasoning involved could hardly be described as 'analytical'. Rather then being merely 'explicative', it appears to be considerably 'augmentative' exactly in Kant's sense.[4] We argue that a similar augmentative process is involved when the natural deduction rules that make use of virtual information namely those which are usually called 'discharge rules' are applied. The reasoning agent who applies these rules has *to make an effort to go (temporarily) beyond the information*, *which is actually given to her*, use some virtual information and then come back. *This stepping out and in again of the given informational space* is what makes informativeness of classical propositional logic so invisible and yet so present.[5]

Hence the virtual state of virtual information seems to lie in the fact that this information although it is not deducible from the premises, has to be necessarily taken temporarily into account in order to obtain the conclusion. In this sense D'Agostino (2013) describes virtual information as "information that is by no means contained in the information carried by the premises of an inference, but is still essentially, *if only temporarily*, involved in obtaining the conclusions" (my emphasis).

We consider the mentioning of the "temporality" together with the reference to the "stepping in and out from the given informational space" of great importance. The first because it is at odds with one of the basic assumptions in logic, i.e. that reasoning doesn't have *in principle* any relation to time. An assumption that more recent considerations on "computational complexity" have already forced us to revise severely, given the distinction between feasible procedures, i.e. procedures

[4] See from the *Critique of the pure reason* (Kant 1787: 33): "Analytical judgements (affirmative) are therefore those in which the connection of the predicate with the subject is cogitated through identity: those in which this connection is cogitated without identity, are called synthetical judgements. The former may be called *explicative*, the letter *augmentative* judgements; because the former add in the predicate nothing to the conception of the subject, but only analyse it into its constituents conceptions, which were thought already in the subject, although in a confused manner; the latter add to our conceptions of the subject a predicate which was not contained in it, and which no analysis could ever have discovered therein." (*my translation*).

[5] My emphasis.

that run in polynomial time, and those which are not. The second one because it underpins the reflecting power of a reasoning person to assess the actually pursued demonstration strategy during its deployment, and redirect it when ineffective, even if it should be formally correct. We understand the "stepping in and out from the given informational space" as the ability to stop just following the rule of mechanically computing on the base of the information given and to go in search for new data patterns that can enrich, even if only "virtually", our informational environment in order to overcome a computational impasse. Hence, to be able to recur to virtual information is of great importance for an effective use of our reasoning competencies, because it enables us to go on searching for unexpected solutions when our mechanical calculus doesn't allow us to reach a positive end. Leaving by side the first aspect concerning the temporality, in what follows I will focus on the second one, aiming at eliciting the conditions, which permit us to access virtual information and take them into account.

12.2.1 Formalizing the Use of Virtual Information Thanks to the Informational Meaning of the Logical Operators

As a paradigmatic example for the use of virtual information in propositional logic (D'Agostino 2013) takes the quite standard strategy one employs to make a non-trivial step of the Sudoku. There are cases where in a given cell n_{ij} we do know with certainty that only one of the two numbers "A" or "B" can be correctly displayed, but given our actual informational state we are not able to find out which of them. Interestingly enough, by trying to insert both numbers alternatively and looking at the disposition of the numbers in the other cells, sometimes we can easily conclude that another cell m_{ij} can be occupied only by a certain number "C", without actually having to decide which number has to occupy the given cell n_{ij}. To infer this conclusion is enough to state that – no matter which one of the two numbers "A" or "B" occupies it – in both cases the cell m_{ij} can only assume the value "C". As the choice between "A" and "B" in n_{ij} remains eventually undecided, but we have to fill it *temporarily and alternatively* with both values, we can regard the use of this information – "charged" during the reasoning but "discharged" by the conclusion – as "virtual". We *step in* an informational environment where we act *as if* the information that n_{ij} is "A", and then n_{ij} is "B" were really at our disposal, we employ this information in our reasoning, and eventually we *step out* from this virtually enriched informational environment and go on ignoring the value of n_{ij}, but have in fact acquired a new informational result, namely that "C" occupies the cell m_{ij}.

D'Agostino (2013) describes it as "the kind of provisional assumptions that occur in the so-called 'discharge rules' of Gentzen's natural deduction and, more generally, in any kind of "reasoning by cases" and formalizes this reasoning introducing "the informational meaning of the logical operators". Differently from

the standard semantic of the Boolean operators, which defines their meaning according to their truth-values, in this semantics the meaning of a logical operator is specified solely in terms of the information that is actually possessed by an agent. Hence the meaning of the logical operators is redefined for signed sentences $T_a(x)$ (respectively $F_a(x)$) asserting that "'*x*' is *true* for an agent *a*" or "*a* knows that '*x*' is *true*" (respectively *false*). In this case we can easily verify that following deduction is correct:

$$\begin{array}{l} T_a\ \varphi \vee \psi \\ T_a\ \varphi \rightarrow \theta \\ T_a\ \psi \rightarrow \theta \\ \hline T_a\ \theta \end{array}$$

Now if the agent assumes that φ is true (that is $T_a\ \varphi$), then by modus ponens she can conclude from the second premise that θ must also be true. On the other hand if the agent assumes that ψ is true (that is $T_a\ \psi$), then by modus ponens she can conclude from the third premise that θ must also be true. Hence although the agent *a* holds neither the information "φ is true" nor the information "ψ", observing that θ necessarily follows from both, by knowing that their disjunction is true, *a* can conclude that θ has to be necessarily true: in the end we have that $T_a\ \theta$ follows from $(T_a\ \varphi \vee \psi)$.

It has nonetheless to be noticed that: (*i*) the information carried by $T_a\ \varphi$ and $T_a\ \psi$ are by no means contained in the premises $T_a\ \theta$ definitely depends on an argument that is *ampliative* since it involves the use of information not actually held by *a*, i.e. the inference is somehow "synthetic" in Kantian sense; and (*ii*) the two pieces of information $T_a\ \varphi$ and $T_a\ \psi$ have been temporarily assumed for the sake of the argument ("charged") and eventually ignored once the conclusion has been reached ("discharged"), hence they can be properly defined "virtual".

Given the use of virtual information in many demonstrations, D'Agostino and Floridi propose a refinement of the concept of analyticity: pure analytic reasoning should be defined, in a negative way, as the one that doesn't need to introduce any information that is not contained in the premises even implicitly, hence it makes no use of what they call "virtual information". And eventually D'Agostino (2014) suggests that according to the increasing use of virtual information in reasoning one can measure the degree of "syntheticity" of a valid inference of classical propositional logic and eventually link this aspect with the corresponding measure of the cognitive effort required to perform that inference. In doing so he refers explicitly to Hintikka's distinction between depth and surface information:

> We suggest that the depth at which 'virtual information' must be invoked in order to recognize the validity of an inference can be taken as a measure of the 'cognitive effort' required to perform this task. From an AI perspective, this cognitive effort is reflected by the computational complexity of the corresponding decision problem, i.e. the problem of deciding whether or not a certain conclusion follows from the premises when virtual information can be used only up to a given fixed depth.

In our view it is precisely this strong connection between "virtual information" and "cognitive effort" which reflects in a rather interesting and fruitful way the Kantian and more in general "transcendental" idea of "synthesis". We suggest that Kant's *transcendental philosophy* offers us an account of the agent's power to "step in and out from an informational environment" and, in doing so, to access to the field of virtual information and make use of it. But for doing so we have to consider how not only the *theoretical reason*, but also the *practical reason* plays a fundamental role in reasoning and how the conjoint use of both is a condition for accessing the space of virtual information. This leads us to a more general consideration concerning the relations between philosophy of information and Kantian or, better, transcendental philosophy. Already Hintikka's approach to the philosophy of information was presented in fact in a Kantian frame, and to support this statement it is worth remembering the full title of his book: *Logic, language games and information. Kantian themes in the philosophy of logic*. The informational meaning of the logical operators proposed by D'Agostino puts this link in a peculiar light.

12.2.2 The Informational Meaning of the Logical Operator as Transcendental Approach to Their Semantics

Before entering the analysis of Kant's philosophy I would suggest that rather than just leading to a solution with a "Kantian flavour", D'Agostino's approach can be seen, more generally, as a *transcendental move in itself*. In the *Critique of the Pure Reason* Kant formulates following definition of "transcendental" – probably the most famous and general one he ever offered: "I entitle transcendental all knowledge which is occupied *not so much with objects as with the mode of our knowledge of objects* in so far as this mode of knowledge is to be possible a priori". (Kant 1787: 43, my emphasis).

I judge D'Agostino's proposal to set the definition of the meaning of the Boolean operators only in terms of the information that we *actually possess* D'Agostino (2013 and 2014) a sort of *transcendental downplay*, if compared to the standard truth conditions, which hinge on the classical information-transcending notions of truth and falsity as primary semantic notions. As in Kant's case the aim of D'Agostino's shifting from an information-transcending (i.e. agent transcendent), to an information-assuming (i.e. agent-immanent or *transcendental*) approach is to rule out some highly problematic metaphysical presuppositions required by the first. The standard view was in fact leading to the assumption of an ideal omniscient agent, which has to be considered a quasi-theological assumption, given the fact

that it cannot be even approximated by any real agent.[6] We remember that Kant's *Critique of the pure reason* (Kant 1787) aimed at defining the boundaries of our reason by rigidly restricting the kind of objects we can make science of, to the ones we can really meet in our experience, i.e. that are suitable to be grasped in space and time. This meant to deny any value to the so-called Rational Theology as we will never make a spatio-temporal experience of God.

But the informational turn in the semantics of the logical operators can be considered Kantian, or *transcendental*, also inasmuch as it casts light on the a-priori, but *actively* deployed contribution of a *subject*, i.e. agent to the reasoning she performs, more than on the pure *objectivity* of the rules her reasoning would *passively* follow. This contribution is moreover concretely reflected in the reasoning strategies actively chosen by the agent. It is, so to say, as if the subjectivist side of the reasoning would get more weight compared to its objective side, leading to a kind of "Copernican Revolution" in the Boolean logic. The very meaning of a logical operator has not to be defined any more according to the "*truth or falsity*" of a sentence as a *value in itself* that we might never be able to access to and therefore have to register as merely given to our mind from the outside; only "*our knowledge of the truth or falsity*", of a sentence i.e. our actual informational status, has now to be taken into account for defining the meaning of the logical operators that it contains. Hence the agent is somehow contributing to define their meaning according to her subjective informational status, i.e. to contents of her mind, which she either actually possesses or at least can retrieve thanks to a feasible, i.e. in the worst-case polynomial, procedure.

In order to underpin the transcendental value of reformulating the meaning of the logical operator in informational terms we paraphrase of Kant's above quoted definition in following D'Agostinian way: I entitle transcendental all knowledge (of the meaning of logical operators) which is occupied not so much with reference to (objective) *truth or falsity of a sentence* as with the (subjective) mode of our *knowledge* (*scil*. information) *of the truth or falsity of a sentence* in so far as this mode of information is to be possible a priori (i.e. actually retrievable). Hence I qualify D'Agostino's strategy as transcendental inasmuch as the definition of the meaning of logical operators is occupied not so much with reference to *truth or falsity* of a sentence as with *our information of the truth or falsity* of a sentence in so far as this information is retrievable. With the effect that the meaning of Boolean operators would be understood as a phenomenon of our informational status, i.e. depending on our knowledge of the truth or falsity of a given sentence.

[6]For a critique to the assumption of the logical Omniscience see D'Agostino (2010) as well as d'Agostino-Floridi (2009).

12.3 What Is Practical Reason for Kant and How It Relates to Virtual Information

12.3.1 Kant's Practical Reason

The very interesting characteristic of the practical reason is that of conveying the freedom of the human being by determining autonomously the law for its will. To act morally means in fact to obey to a self-prescribed law without being forced by any external, i.e. physical constraint and basically against any inclination, all of which are oriented at satisfying our desires. Kant names this capability: "self-legislation of the practical reason".

Some years before, Kant, in the *Groundwork for the Metaphysic of Morals*, linked more explicitly the manifestation of freedom in the moral behaviour to the fact that human actions are not merely ruled by "laws of nature [...] in accordance with which everything happens" but also by "laws in accordance to which everything *ought to* happen" (Kant, 1785: 3, *my emphasis*). It seems hence that the moral law is based on the human possibility of thinking in terms of "ought to" as an alternative to the "must" which expresses the way natural laws rules. Hence the foundation of the moral is aimed at the explanation of the force that the moral law, i.e. a law formulated on the basis of a mere "ought to", can set on the human behaviour so as to determine it with the same necessity as if it was submitted to a law of nature. Accordingly Kant states: "We are not talking here about whether this or that happens, but rather reason commands, for itself and independently of all appearances, what ought to happen" (Kant, 1785: 24).

As a last point in this general summary of Kant's moral theory, we want just to stress that this focus on the "ought to" as the way the human freedom makes its appearance, is independent from, if not at odds with our empirical experience: our knowledge of what ought to be done, is absolutely independent from the way things are or have been done.

All human beings think of themselves, regarding the will, as free. Hence all judgments about actions come as if they *ought to have happened even if they have not happened*. Yet this freedom is no experiential concept, and also cannot be one, because freedom always remains even though experience shows the opposite of those requirements that are represented as necessary under the presupposition of freedom. (Kant, 1785: 71).

12.3.2 Practical Reason and Virtual Information

Now, if we go back to the example of the Sudoku, we can easily verify that our first conclusion that in cell n_{ij} can be only entered one of the two numbers "x" or "y" is for sure the result of the use of the theoretical reason. In fact we attain this information by just looking at the disposition of the numbers in the cells and by

mechanically applying the rule of the game that each row, column and square cannot contain the same number more then once. At this point however our theoretical reason gets stuck, as we cannot, by simply following this rule, fill with certainty any other cell.

Hence, in order to make use of the information consisting in the above-mentioned alternative as a premise for going on filling other cells, we firstly have to be *willing* that the given cell n_{ij} is alternatively occupied with the number "x" and then with the number "y" and then consider which outcomes this move has for other cells. What we are striving to attain in the end is a pure theoretical result, i.e. to enhance our informational status by filling a further cell m_{ij}; but in order to do this we cannot rely on the theoretical reason alone, as this one has already come to an end. In fact, insofar as our informational status doesn't force us to insert any of the two numbers in the cell n_{ij} – and therefore we wouldn't even really be entitled to do so – while we are inserting alternatively the two numbers we are not *naturally* necessitated to do it, but we are somehow *morally* forcing us to perform this attempt. In doing so, we are of course still making use of the games rule, but we don't let it run for fixing the value of *a* cell, as usually done, rather for testing the outcomes of the given alternatives in the expectation that they give us useful information for other cells. So once we act like this we are not, properly speaking, simply applying the rule, i.e. passively following it, but we are rather actively interpreting the way to apply it and hence let the rule be ruled by a law leading it from above.

The question now is: which faculty is able to prescribe us laws and what kind of law we do follow? My answer is that we are making use of the faculty of our reason to self-prescribe laws to itself, and these are formulated in form of an "*ought to*" or a "*should*". So, when we are missing a theoretical obligation, when the mechanical use of a rule blocks us in the reasoning process, we can still recur to the self-legislation of the "*ought to/should*" to overcome our theoretical impasses. The latter is usually displayed in the moral, but it can also do its work as a support for the theoretical reason.

Hence I suggest here that precisely this *practical use of our reason* opens up the space of *virtual information* to our theoretical reason and allows us to *freely* step in and out of it, i.e. to take virtual information into account, charging them in the reasoning and discharging them in the conclusions. Virtual information are used by the theoretical reason, but set at its disposal by the practical one. Reasoning somehow independently, if not against the mere application of the rule imposed by the game, can only be done thanks to another legislative power that prescribes to our reasoning a new rule, following Kant's idea that "reason commands, for itself and independently of all appearances, what ought to happen" (Kant 1785: 24).

12.4 Conclusion and Remarks

If this conclusion holds and we follow D'Agostino's idea, that the difference between *analytic* and *synthetic a-priori* reasoning is owed to the agent's recourse to virtual information, we could also conclude that synthesis a-priori is the result of a

combined use of theoretical and practical reason. And as synthesis means to increase our information, we could also affirm that the actual, i.e. not merely psychological, enriching of our informational status during the process of demonstrating a theorem is due to the fact that the demonstration is led by a strategy implying the conjoint use of theoretical *intellect* and *will*.

The meaning of this remark would be that the "informational turn" proposed by D'Agostino in logic would also imply a sort of "practical turn" in epistemology. In fact rationality would be here regarded as the faculty of an agent to process information in order to take decisions, i.e. willing to act according to the output of the reasoning. Hence it's not surprising that a stronger focus on the informational meaning of our knowledge let also the practical component of our rationality come explicitly to the foreground.

Now, according to D'Agostino's results, it seems that among the things someone can do with the information she possesses there is also their use for acquiring further information. This is definitely at odds with the idea that data-processing can only be analytical and in fact we saw at least one way of processing some information we possess that enriches our informational status. Hence one of the meanings of D'Agostino's proposal to restate the existence of the synthesis a-priori is to confirm the possibility to increase our information without looking for new empirical data, but just operating "smartly" with the information we already have. Our suggestion is that this happens when the processing effort does overcome the mere calculation and employs reasoning strategies leaded by the practical reason.

But it is not only that Kant's theory helps us to shed a light on the way we handle virtual information; the established link between virtual information and practical reason helps us also vice versa to better focus on the condition of possibility of the Kantian synthesis a priori more in general. Namely if already at the basic level of the propositional logic we can speak of synthesis a priori when virtual information enter the argument, but this actually depends on the practical use of reason for the sake of theoretical purposes, then the synthesis a priori should always be depending on the practical use of the reason even without leading to any action. And actually this cooperation on the field of our knowledge goes hand in hand with the parallel cooperation between theoretical and practical reason in the agency.

References

Abbott, Edwin A. 1884. *Flatland: A romance of many dimensions*. London: Seeley & Co.

Carnap, R., and Y. Bar-Hillel. 1953. An outline of a theory of semantic information. In *Language and information: Selected essays on their theory and application*, ed. Y. Bar-Hillel, 221–274. Reading: Addison-Wesley.

D'Agostino, Marcello. 2010. Tractable depth-bounded logics and the problem of logical omniscience. In *Probability, uncertainty and rationality*, ed. F. Montagna and H. Hosni, 245–275. Springer.

D'Agostino, Marcello. 2013. Semantic information and the trivialization of logic: Floridi on the scandal of deduction. In *Information 2013*, vol. 4, 33–59.

———. 2014. Analytic inference and the informational meaning of the logical operators. *Logique et Analyse* 227: 407–437.

D'Agostino, Marcello, and Luciano Floridi. 2009. The enduring scandal of deduction. Is propositionally logic really uninformative? *Synthese* 167: 271–315.

Floridi, Luciano. 2011. *The philosophy of information*. Oxford: Oxford University Press.

Hempel, Carl Gustav. 1945. Geometry and empirical science. *American Mathematical Monthly* 52: 7–17.

Hintikka, Jaakko. 1973. *Logic, language games and information. Kantian themes in the philosophy of logic*. Oxford: Clarendon Press.

Kant, Immanuel. 1785. *Groundwork for the metaphysic of the moral*. Ed. and Trans. Allen W. Wood. New Haven/London: Yale University Press, 2002.

———. 1787. *Kritik der reinen Vernunft*, in: *Kant's Gesammelte Schriften* ‚Akademieausgabe", Königlich Preußische Akademie der Wissenschaften, Berlin 1900ff, III Band, 1904, 1911.

———. 1789. *Kritik der praktischen Vernunft*, in: *Kant's Gesammelte Schriften* ‚Akademieausgabe", Königlich Preußische Akademie der Wissenschaften, Berlin 1900ff, V Band, 1913.

Chapter 13
A Kantian Cognitive Architecture

Richard Evans

Abstract In this paper, I reinterpret Kant's *Transcendental Analytic* as a description of a cognitive architecture. I describe a computer implementation of this architecture, and show how it has been applied to two unsupervised learning tasks. The resulting program is very data efficient, able to learn from a tiny handful of examples. I show how the program achieves data-efficiency: the constraints described in the *Analytic of Principles* are reinterpreted as strong prior knowledge, constraining the set of possible solutions.

Keywords Kant · Critique of pure reason · Rule induction · Unsupervised learning · Data efficiency · Cognitive architecture · Computational modeling · Original intentionality · Cognitive agency

13.1 Introduction

In this paper, I shall reinterpret part of Kant's *Critique of Pure Reason* as a specification of a cognitive architecture. I will describe a computer implementation of this architecture, and show how this program has been applied to an open problem in AI.

Now this project may seem, on the face of it, absurd: why should a book written in the eighteenth century have anything to teach us now? I will argue that this is not as unpromising as it might, at first, appear. Kant still has something to teach us. His insights have not yet been fully absorbed into cognitive science or AI.

I shall describe two Kant-inspired computer programs that are able to perform unsupervised learning from a tiny handful of examples. Now the ability to learn

R. Evans (✉)
Imperial College London, London, UK

DeepMind, London, UK
e-mail: RichardPrideauxEvans@imperial.ac.uk

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_13

from a handful of data, often called *data-efficiency*, requires strong prior knowledge. But if strong priors encode *domain-specific* information, then our machine-learning system is parochial, tied to a particular subject matter. What we really want, if only we could get it, is a set of strong priors that are also *domain-independent*.

In this paper, I shall show how the constraints described in Kant's *Analytic of Principles* can be reinterpreted as a set of strong, domain-independent priors: a set of constraints on any agent that is trying to turn its sensory input into a coherent unified experience of an external world. First, in Sect. 13.2, I shall pick out a particular argumentative strand that runs through the first half of the First Critique. I shall describe a set of conditions that, Kant holds, must be true of any agent that is able to make sense of its sensory data. Next, in Sect. 13.3, I shall reinterpret these Kantian claims as a specification of a cognitive architecture. The Kantian constraints are translated into constraints on the types of rules generated by a rule induction system. Finally, in Sect. 13.4, I shall describe two applications of this architecture. In one, the agent is placed in a simple two-dimensional grid-world, and must make sense of the sensory data he receives. In the second application, the agent's sensory input is a one-dimensional string of symbols. Making sense of these strings of symbols amounts to solving a standard verbal reasoning task. Surprisingly, the Kantian agent is able to achieve human-level performance in this verbal reasoning task, with no prior training data and no hand-engineered feature recognition.

13.2 Original Intentionality Via Synthetic Unity

I define a **sensory agent** as some sort of animal or device, equipped with sensors, whose actions depend on the state of its sensors. It might have a temperature gauge, a camera with limited resolution, or a sonar that can detect distance. The sensory agent is continually performing what roboticists call the *sense-act cycle*: it detects changes to its sensors, and responds with bodily movements.

A thermostat, for example, is a simple sensory agent. When it notices that the temperature has got too low, it responds by increasing the temperature. Now although the thermostat has a sense-act cycle, it does not experience[1] the world it is responding to. *We* count the perturbations of its gauge as representations of the temperature in the room it is in, but *it* does not. The gauge movements count as temperature representations *for us*, but not *for the thermostat*. Nothing counts as anything for the thermostat. It just responds *blindly*.

The thermostat does not, in other words, have *original intentionality* (Haugeland 1990). *We* might interpret some of its activities as representations, but *it* does not.

[1] In making this claim, I am assuming a suitably red-blooded notion of "experience". Of course, for some sufficiently thin notion of "experience", the thermostat must "experience" the world in order to act at all. But there is a difference between merely responding to a stimulus and *making sense* of that stimulus: reinterpreting the stimulus as a representation of a coherent external world. The latter is "experience" in the strong sense I am using it.

We can distinguish between derivative and original intentionality using the activity of counting-as[2]:

- x has derivative intentionality in representing p if an agent y (distinct from x) counts x's activity as x's representing p
- x has original intentionality in representing p if x himself counts x's activity as x's representing p

What distinguishes an agent with original intentionality, a **cognitive agent**, from a mere sensory agent is that the former *counts its own sensings as* representations of a determinate external world. It interprets its own sensory perturbations as a representation of a coherent unified world of external objects, interacting with each other.

One of Kant's fundamental questions is:

> *What does a sensory agent have to do, in order for it to count its own sensory perturbations as experience, as a representation of an external world?*

What, in other words, must a sensory agent do to be a cognitive agent?

Note that this is a question about intentionality – not about knowledge. Kant's question is very different from the standard epistemological question:

> Given a set of beliefs, what else has to be true of him for us to count his beliefs as knowledge?

Kant's question is *pre-epistemological*: he does not assume the agent is "given" a set of beliefs. Instead, we see his beliefs as an *achievement* that cannot be taken for granted, but has to be *explained*:

> Understanding belongs to all experience and its possibility, and the first thing that it does for this is not to make the representation of the objects distinct, *but rather to make the representation of an object possible at all* (Kant 1781)(A199, B244-5)[3]

Kant asks for the conditions that must be satisfied for the agent to have any possible cognition (true *or* false) (Kant 1781)(A158, B197). Kant's question, in the first person, is:

> *What do I have to do, in order to count these sensory perturbations as my experience?*

His answer, roughly, is:

> I count this plurality of sensings as my experience if I combine them together in the right way

[2]Note that I am not *defining* intentionality in terms of the activity of counting-as (which would be uninformative). Rather, I am using counting-as to *distinguish* between original and derivative intentionality. Later, counting-as will itself be explicated in terms of the construction and application of rules.

[3]All such references [A, B] are to the A and B editions of the *Critique of Pure Reason*, (Kant 1781).

What, then, does Kant mean by "combine", and what does he mean by "the right way"?

First, in Sect. 13.2.1, I will describe what Kant means by "combine". To anticipate, there are two types of combination, achieved by applying two types of rules (rules of composition and rules of connection). Second, in Sect. 13.2.2, I will describe what Kant means by the "right way". To anticipate, combining in the right way means connecting the cognitions together via certain relations, so that the plurality of cognitions becomes a totality connected in time. The constraints described in the *Analytic of Principles* are constraints on the construction and application of rules that Kant claims are severally necessary and jointly sufficient for an agent to construct a coherent representation. In Sect. 13.2.3, I will outline the general argument structure that underlines each of the four *Principles*.

13.2.1 The Basic Activity of Combination

The activity at the heart of Kant's theory is the mental act of **combination**, of bringing cognitions together, "running through and holding together this manifoldness" (Kant 1781)(A99). Kant explains what he means by "combination" in Kant (1781)(B201n): I combine a plurality of cognitions together when I subsume then under a "mark". Kant says little about what a "mark" is, given its load-bearing role in his theory. *"Merkmal"* is typically translated as "mark", but it can also be translated as "feature". Kant's mark is not a shared linguistic symbol. It is rather what computer scientists call a "gen-sym": a generated symbol, an atomic identifier. When it is first created, a mark is just an uninterpreted symbol. But by constructing inferential rules that relate this mark to others, the agent can elevate it into a concept.

Combining, then, is subsuming cognitions under a mark. For example: if this configuration of sensors is turned on, then I count their being-on as representing a nose. Or: if this pattern of sensors counts as representing a nose, and this other pattern counts as representing an eye, then the aggregate pattern of sensors counts as representing a face.

13.2.1.1 Combination Can Only Be Performed Indirectly Via the Construction and Application of Rules

Although this combining activity is fundamental, it cannot, according to Kant, be performed directly by the agent. The agent cannot just bring representations together *willy-nilly*.[4] Combining is not something he can *just do*. On the contrary, the *only way*, according to Kant, that the agent can perform the activity of combination

[4]By "willy-nilly", I mean without justification from the application of a rule. Kant's view is that the only mental actions that are justified are actions that result from applying a rule. What leaves room in this stern vision for spontaneity and autonomy is that the rules are not imposed from outside; rather, they are self-legislated.

is indirectly, by *applying general rules*[5] *that it has constructed*. This is Kant's surprising claim.

In a revealing footnote [B201n], Kant distinguishes between two types of rule of combination. **Rules of composition** are rules for combining parts into wholes, producing a part-whole graph united under one element: the totality. A rule of composition produces, if it applies, a defeasible *permission*[6] for the agent to group intuitions together under a mark. For example, if you count this group of sensings as representing an ear, and this group of sensings as representing a nose, then you may count this aggregate group of sensings as representing a face. Whether or not the rule-following agent makes use of this permission will depend on his concomitant commitments.

Rules of composition are described by *defeasible* conditionals. Wittgenstein stresses the defeasibility of such conditionals when discussing what counts as a friendly face:

> When we notice the friendly expression of a face, our attention, our gaze, is drawn to a particular feature in the face, the 'friendly eyes', or the 'friendly mouth etc. . . . *It is true that other traits in this face could take away the friendly character of this eye*, and yet in this face it is the eye which is the outstanding friendly feature (Wittgenstein 1958) (p.145–146)

This is defeasibility in action: in this situation, the features of this eye counts as his having a friendly face; but in another situation, the very same features plus some other additional facial features might count as something entirely different – mocking cruelty, for instance.

The second type of rule is a necessary rule that must be applied when it can be applied. **Rules of connection** produce *obligations* to group representations under a mark.[7] So, for example, if we count this structure as a nose, then we *must* also count it as a facial part – and if we count it as a nose, then we *must not* count it as an ear.

Kant's striking claim is that the mental act of combination is not a self-sufficient action, something you can *just do* – rather, you can only do it by applying these two types of rules. We are used to thinking of *social* activity as constituted: moving your knight to king's bishop three is something you can only do indirectly *by doing something else* – by pushing a wooden object in a certain direction. Similarly, requesting Bob to shut the door is not something you can *just do*: you can only do it by doing something else – perhaps by uttering a sequence of sounds, or by pointing at the door; there are an infinite number of different actions that could constitute

[5]Please note that these Kantian rules do not have to be linguistically articulated or consciously accessible. Rather, the rules that determine the activities of mental combination are implicit and consciously inaccessible, in the same way that the rules of a compiled Prolog program are inaccessible to the executing process.

[6]See Kant (1781)(B201n): "the synthesis of a manifold of what does *not necessarily* belong to each other".

[7]See Kant (1781)(B201n): "the second combination is the synthesis of that which is manifold insofar as they *necessarily* belong to one another".

such a request, but you have to do *one* of them – requesting is not something primitive you can do on your own. But we are not so used to thinking of fundamental *mental* activity as similarly constituted.

All the agent can do is *construct* general rules of the above form, permitting or obligating him to combine representations in a certain way, and then *apply* these rules, thus *indirectly* performing combinations via the construction and application of rules. This claim appears throughout the First Critique[8]:

> everything (that can even come before us as an object) necessarily stands under rules, since, without such rules, appearances could never amount to cognition of an object [B198, A159]

Why can't a cognitive agent perform the activity of combination *directly*, without needing to construct and then apply a rule? We will see why this is so in Sect. 13.2.2. The basic reason, to anticipate, is that combining without rules would not satisfy the condition of *unification* at the heart of K's theory. The unification condition is a set of constraints on the construction and application of *rules*, and so can *only* be applied to a rule-following and rule-constructing agent. Arbitrary combination of cognitions that was unguided by rules would not produce a unity of experience that I could call *mine*; instead, the combined representations would be a "mere play", "less even than a dream" (Kant 1781)(A112). If I could combine representations into intuitions without rules, then there would be no *self* to have the intuitions.

The Kantian rule-following agent is continually constructing the very software that it will then execute.[9] It is always constructing rules, and then interpreting those rules. In fact, the *only* way that it can perceive *anything* is by applying rules it has already constructed in order to make sense of the incoming barrage of sensations.[10]

The Kantian rule-follower, then, is a norm-giving agent who solemnly sets down rules that he will then obediently follow. He only allows himself to perform acts of mental combination if these acts are shown to be permitted by rules he has previously constructed.

[8] See also [A105], [A177, B220].

[9] In computational terms, think of a meta-interpreter that is able to construct pieces of code as *data*, and then execute these new pieces of code.

[10] Kant makes the same point in the *Metaphysical Deduction*: "The same function that gives unity to the different representations *in a judgement* also gives unity to the mere synthesis of different representations *in an intuition*, which, expressed generally, is called the pure concept of the understanding. The same understanding, therefore, *and indeed by means of the very same actions* through which it brings the logical form of a judgement into concepts by means of the analytical unity, also brings a transcendental content into its representations by means of the synthetic unity of the manifold" (Kant 1781)(A79, B104-5). In other words, there is only *one* process (a process of constructing and applying rules) which explains *both* how we form judgements *and* how we form intuitions.

13.2.1.2 Constructing and Applying Rules

The rule-following agent can perform two types of activity: he can construct a rule, and he can apply a rule he has already constructed. Kant says it is the job of the faculty of understanding to construct rules:

> Sensibility gives us forms (of intuition), but the understanding gives us rules. It is always busy poring through the appearances with the aim of finding some sort of rule in them. (Kant 1781)(A126)

Recall that there are two types of rule (rules of composition, and rules of connection), so there are two types of rule construction. Constructing rules of composition is forming perceptual rules, rules of apprehension for counting particular configurations as parts of objects. For example, the agent adds a new rule that, if some of its sensors satisfy such and such a condition, it may count them as representing an ear.

Constructing rules of connection is forming concepts or making judgements. Forming a concept is constructing a set of rules that describe the inferential connections between this concept and others. So, for example, to form the concept of "tree", we need rules of composition for saying under what conditions a set of sensury perturbations count as representing a tree. But we also need rules of connection for linking this concept with others. For example, if we count it as a tree, we must also count it as a plant.[11]

Making a judgement is also constructing a rule of connection. If we form the judgement that "All men are mortal", this is just to adopt the rule of connection: if I count a cognition as a man, then I must also count it as mortal. But this inferential understanding of judgement applies to categorical statements just as much as to hypothetical statements. To form the judgement that "Caesar is a general" just is to adopt the rule: if I count a cognition as Caesar, then I must also count it as a general (Longuenesse 1998).

This is why Kant says (Kant 1781(A126)) that the faculty of constructing rules is also the faculty of concept-formation and judging: both concept-formation and judging are just special cases of the more general ability to construct rules.

Next, I shall turn to the process of *applying* the rules that the understanding has constructed. If the rule applies in a particular situation, a norm is operative: either the agent must combine the representations under a certain mark (if the rule is a rule of connection), or it may do so (if it is a rule of composition). If it is a rule of composition, then all the agent knows is that he *may* perform the combination activity – he does not *have* to do so. Consider, for example, Jastrow's famous duck-rabbit (Fig. 13.1). Focus on the lines on the left of the image. Now we have two rules

[11] Some of the connection rules involved in characterising a concept do more than simply state that one concept is a sub-concept of another, or that one concept excludes another. Some of them relate the concept to another concept only conditionally – dependent on the existence of external factors. For example: "If the weather gets cold, trees lose their leaves", "If a tree gets no water, it perishes" (Longuenesse 1998). Some of the conceptual inference rules, in Kantian terms, are hypothetical rather than categorical.

Fig. 13.1 Jastrow's duck-rabbit

of composition that apply to these lines: we can count these lines as a mouth, or as a pair of ears. Now there is a rule of connection that prevents us from applying both: if something is a mouth, then it is not a pair of ears. We may apply either rule of composition – but we must not apply both. What makes us decide which to apply?

Kant argues convincingly that it cannot be a further *rule* that tells us which to apply. If we needed rules to determine which rules to apply, then those determining rules would themselves need further rules to determine their application, and so on, generating a vicious regress (Kant 1781)(A133, B172).

Kant defines the *imagination* as the faculty responsible for **applying** the rules that the understanding has constructed. As the duck-rabbit picture shows, the imagination has some *choice* about how to apply the rules of composition (Kant 1781)(B151). This is why Kant says that both understanding and imagination involve *spontaneity* – the understanding has a choice about which rules to construct; and then, once it has constructed them, the imagination has a further choice about which rules of composition to apply:

> It is one and the same spontaneity that, there under the name of imagination and here under the name of understanding, brings combination into the manifold of intuition (Kant 1781)(B162n)

Note that it is only when applying rules of *composition* that the imagination has a choice about which to apply. When it comes to applying rules of *connection*, the rule-following agent is *obligated* to perform the required mental activity.

To summarise, a rule-following agent is a type of sensory agent who can combine representations by constructing and applying rules. Given that there are two types of activity (constructing and applying rules), and two types of rule (rules of composition and rules of connection), we have a square of operations.

	Rules of composition	Rules of connection
Constructing	Forming perceptual rules	Forming concepts and judgements
Applying	Forming intuitions	Inferring properties of objects

Recall our original question:

What must I do, in order to count these sensory perturbations as my experience?

The rule-following agent is a central part of Kant's answer:

- A sensory agent is a cognitive agent if he counts his sensings as representing an external world
- He counts these sensings as representing an external world if he combines those sensings together in the right way
- He combines his sensings together in the right way if he constructs and applies a set of rules that satisfy a set of (as yet unspecified) *constraints*

The next question, then, is: *what set of constraints on the construction and application of rules are severally necessary and jointly sufficient for counting this plurality of sensory perturbations as representing an external world?* The next section will describe the constraints involved.

13.2.2 Combining in the "Right Way"

The constraints on the activity of combination are specified in the *Analytic of Principles*. The argument justifying these particular constraints is spread through the *Transcendental Deduction*, the *Schematism*, and the *Principles*:

1. Counting this plurality of sensory perturbations as *my experience* requires *connecting* the representations together
2. Connecting the representations together requires bringing the representations under a *relation*
3. The only medium that can connect all my representations is *time*
4. Connectedness in time involves four activities[12]:
   (a) constructing moments in time
   (b) filling time
   (c) ordering time
   (d) constructing the totality of time

I shall go through these points in turn.

13.2.2.1 If I am to Make them Mine, I Must Connect them Together

For the representations to be mine, I must *make* them united:

> Combination does not lie in the objects, however, and cannot as it were be borrowed from them through perception and by that means first taken up into the understanding, but is rather only an *operation* of the understanding (Kant 1781)(B134-5)

[12]These activities are described in Kant (1781)(B185, A146) .

Nobody is going to unify my sensings for me – it is an *activity I must perform myself* if I am to experience anything at all. I must *work* to bring my representations together into a unity. Kant calls this requirement the "supreme principle of human cognition".

13.2.2.2 Connecting Cognitions Together Requires Uniting Them Under Relations

The only way we can connect cognitions together is by placing them under relations. Connecting the cognitions via relations is the only way "to make them fit for a thoroughgoing *connection* in one experience" (Kant 1781)(B185, A146).

13.2.2.3 The Only Medium That Can Connect All My Representations Is Time

Our representations are connected by bringing them together under certain relations. Some of my representations are intuitions about the external world, and others are thoughts about my own inner states (beliefs, pain, etc). Now although I can place my outer intuitions in space, my inner intuitions are not spatially-located. My inner intuitions are ordered in time, but not in space. The only medium in which I can place *all* my intuitions is time:

> There is only one totality in which all of our representations are contained, namely inner sense and its a priori form, time. (Kant 1781)(B194, B155)

So the *only* marks that can relate *all* my representations are relations involving time:

> Time is the formal condition of the manifold of inner sense, thus of the connection of all representations (Kant 1781)(A138, B177)

13.2.2.4 Connecting my Representations in Time Requires Four Activities of Time-Determination

Kant lists four activities of **time-determination** which he claims are necessary and sufficient for unifying all representations in time (Kant 1781)(B184, A145). The first activity is *constructing moments in time*: constructing successive apprehensions[13] of an object, and representing a moment in time as a collection of simultaneous apprehensions of objects, apprehensions that are organised in a part-whole totality. The second aspect of time-determination is *filling time*: determining cognitions with sufficient fine-grainedness that any intermediate moment in time can be constructed.

[13] I use the Kantian term **apprehension** to denote a time-slice of an enduring object at a particular moment in time. Throughout, I use "apprehension" and "object-slice" interchangeably.

Between any two moments, the agent can construct an intermediate moment. The third aspect is *ordering time*: placing moments of time in a determinate order, a total ordering on moments. The fourth aspect of time-determination is *constructing the totality of time*: specifying which of the various candidate combinations of representations are coherent moments in time, and which are impossible.

13.2.3 The General Structure of the Four Principles

The constraints described in the *Analytic of Principles* are constraints on the construction and application of rules needed to satisfy the four activities of time-determination above.

First, I shall lay out the common argument structure behind each principle before addressing them individually.

Let us distinguish, following Sellars (1968), between:

- (A) **representeds**: the things that are represented
- (B) **representings**: activities that are *of*, or *about* representeds
- (C) **rules**: general procedures that apply in many situations; when a rule is applied, it results in a representing activity

There are various connections between these three elements. First, the representeds (A) are the content of the representing activities (B): representings (B) are of, or about representeds (A). We are only able to have representeds (A) by doing representings (B). Second, we can only perform representings (B) by applying rules (C). It is only by applying rules (C) that we are able to perform (B).

Corresponding to these three types of construct are three types of constraint:

- **A-Constraint**: constraints on representeds: constraints on the content of the representing activity
- **B-Constraint**: constraints on representings: constraints on the activity of representing itself
- **C-Constraint**: constraints on the rules that are constructed and applied

The basic structure of the argument in each principle is:

1. start with A-constraints: constraints on the representeds
2. derive from the original A-constraints further A-constraints, using the additional premise that time is not directly perceived
3. move to B-constraints: constraints on the representings, on the activity of combination
4. move to C-constraints: constraints on the rules used to produce the activity of combination
5. finally, derive from the C-constraints additional necessary A-consequences: things that must be true of any representation that has achieved unity (Fig. 13.2)

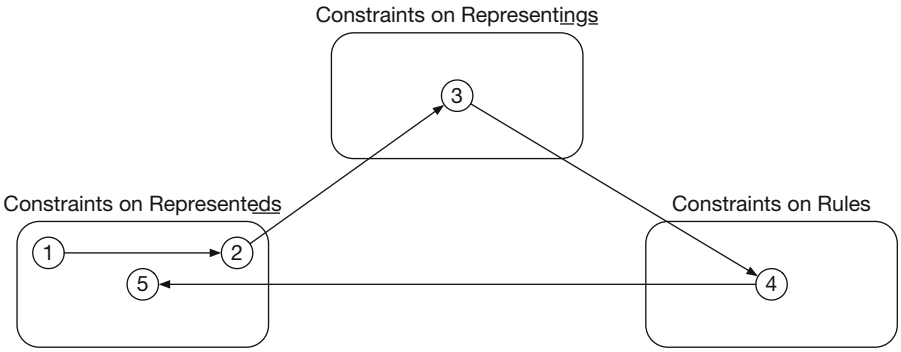

Fig. 13.2 The basic argument structure

13.2.3.1 Constraint 1: The Representeds Must Achieve a Certain Aspect of Time Determination

We start with a requirement on the representeds (A-Constraint): that the cognitions must satisfy an aspect of time-determination in order to achieve unity. Recall from Sect. 13.2.2.4 that there are four aspects of time-unity described in the *Schematism*: construction of moments, filling time, ordering time, the sum total of time.

We ask: what must be true for this group of representeds to achieve this aspect of time determination?

13.2.3.2 Constraint 2: There Must Be a Certain Relation That, If Satisfied by the Representeds, Allows the Representeds to Achieve a Certain Time Determination

In Constraint 2, we remain focused on A-constraints. We just move from a general constraint to a more specific constraint. In this phase, we use the additional premise that *time cannot be directly perceived*. Given that time cannot be directly perceived, the only way we can achieve the general high-level constraint is by the existence of a *relation* satisfying various properties.

We ask: what must be true of this plurality of representeds for there to be a relation satisfying various properties?

13.2.3.3 Constraint 3: The Activity of Combination (i.e. Rule-Application) Must Satisfy a Certain Property

In Constraint 3, we move from an A-constraint (on representeds) to a B-constraint (on the activity of representing). Recall that the representings are just acts of combination, and these acts of combination are just applications of rules. We ask:

what must be true of this *rule-application activity* for the produced representeds to satisfy the relation in Constraint 2?

13.2.3.4 Constraint 4: The Rules Must Satisfy a Certain Property

Now, given that the activity of rule application depends on the types of rules that were constructed, we move in Constraint 4 from a B-constraint on the activity of rule-application to a C-constraint on the types of rules that were constructed in the first place.

We ask: what properties must the rules satisfy if, when they are applied, the rule-application activity satisfies Constraint 3?

13.2.3.5 Constraint 5: Therefore, the Representeds Will Satisfy Various Additional Properties

Now, when we apply rules satisfying certain properties, the products of this rule application (i.e. the representeds) will necessarily have certain properties.

In this phase, we infer from constraints on (C) to additional necessary properties of (A). If we are interested in cognitive science, we are most interested in Kant's derivation of Constraints 1–4. If we are interested in metaphysics (specifically, Kant's argument for various synthetic a priori propositions) then we are interested in the final conclusion, Constraint 5.

13.2.3.6 An Analogy

Suppose you are leading a kindergarten dance class and you need, at some point during the performance, to get the group of five year olds to form a circle. This is the initial general Constraint 1.

Now suppose we are not permitted to draw a circle on the ground. The stage-manager is very strict, and does not permit any tampering with the stage. So the children must be encouraged to form a circle without any external markings to help them. What we want is a relation "right-of" between agents such that, for every agent, there is exactly one agent who is "right-of" him. Furthermore, we require that the transitive closure of "right-of" connects every two agents (Note: of course, this isn't strictly-speaking necessarily a circle. It might be a rather squishy shape. But it will be a closed shape). This is Constraint 2.

Now how do we get the five year olds to achieve this relation? The dance teacher will not be on stage with the children when the time comes for the children to form a circle. We can only get them to form a circle by giving them, in advance, a rule to apply. This is Constraint 3: we have moved from a constraint on the positions of their bodies to a constraint on their rule application.

Now what sort of rule would be suitable? One possible rule, for instance, is: when it is time to form a circle, find somebody whose left hand is free, and hold their left hand with your right hand. This is Constraint 4, a condition on the rules that generate the activity.

Now the conclusion (Constraint 5) is a derived fact about the physical orientation of the dancers, assuming they are following the rules correctly: nobody has a free hand, as each hand is holding some other hand.

This analogy may or may not be helpful, but there is one point of disanalogy that must be mentioned. In Constraint 4, the proposed rule is only one amongst many ways of achieving the right-of relation. In Kant's argument, by contrast, the fourth constraint on the construction of rules is *necessary*.

13.3 Reinterpreting Kant's Theory as a Specification of a Cognitive Architecture

To recap, Kant sees a cognitive agent as a particular type of rule-follower. The only way it can make sense of its sensory given is by applying rules it has already constructed. It is "always busy poring through the appearances with the aim of finding some sort of rule in them" (Kant 1781)(A126). Recall that there are two types of rule: rules of composition are defeasible rules that bring intuitions together; rules of connection are strict rules that operate by necessity, subsuming intuitions under marks. The Kantian cognitive agent must find a set of rules that, when applied to the sensory given, produce a set of representations that satisfy the constraints described in the *Analytic of Principles*. If the results of the rule application process do indeed satisfy the four *Principles*, then the resulting cognitions achieve a unity: they are unified in the medium of time. This is what it is for the agent to achieve original intentionality.

The central idea of this paper is to reinterpret Kant's cognitive agent as a rule-induction system, searching over the space of non-monotonic logic programs. Rules of composition are interpreted as defeasible rules in a non-monotonic logic, rules of connection are interpreted as strict rules. Given a set of rules (i.e. a non-monotonic logic program), a coherent interpretation is a stable model of that logic program. Each of the four *Principles* is interpreted by two constraints: a relational constraint on the stable models that are deemed acceptable, and a structural constraint on the forms of logical rules that are allowed. Making sense of one's sensor readings, according to this interpretation, is searching through the space of non-monotonic logic programs (restricted by the structural constraints) for a program that, when applied to the sensory given, produces at least one model that satisfies the relational constraints.

In the rest of this section, I shall try to explain this central idea in more detail.

13.3.1 Representing Kant's Rules in a Non-monotonic Logic

Consider a logic that contains both strict and defeasible rules. As well as strict rules of the form:

$$h \leftarrow b_1, \ldots, b_n$$

we also have defeasible rules of the form:

$$h \leftsquigarrow b_1, \ldots, b_n$$

The latter means that if $b_1, \ldots, b_n$ all hold, then you may conclude that h, as long as h is compatible with your other commitments. This defeasible implication can be translated away using negation-as-failure[14] and classical negation:

$$h \leftarrow b_1, \ldots, b_n, \mathsf{not}\ \neg h$$

Logics with defeasible rules, defaults,[15] or negation-as-failure, are **non-monotonic** in that $A \models p$ does not imply $A \cup B \models p$.

Defeasible rules are used to model Kant's rules of composition. Strict rules (rules containing no negation-as-failure) are used to model Kant's rules of connection.

13.3.2 Representing Apprehensions (Object-Slices) as Logical Terms

When representing objects that change over time, most work in knowledge representation assumes that the logical terms represent enduring objects, persisting over time.[16] To capture the fact that objects' properties change over time, the predicates are given an additional argument, to represent the time-index, or situation, at which the property holds.

In this model, by contrast, logical terms represent *apprehensions*: slices of objects *at a particular time*. Rules of composition (defeasible rules) produce new terms representing object-slices at a particular moment. The important thing about starting with object-slices and building up to objects (rather than starting with objects, and dividing them into slices) is that *it allows us to focus on the non-trivial task of reidentifying the same object over time*: is this object-slice which I construct from sensor-34 at time t_1 the same object as the intuition which I construct

[14] Here I assume the stable model semantics (Gelfond and Lifschitz 1988) for negation-as-failure.

[15] See Reiter (1980).

[16] For influential examples, see Kowalski and Sergot (1989) and McCarthy (1963).

from sensors 78 and 102 at time t_2? As Frege has emphasised, this question is and should be non-obvious. The reason for starting with object-slices and building up to enduring objects is that it makes explicit the achievement involved in answering this question of reidentification.

13.3.3 Translating Each Principle into a Relational and Structural Constraint

Recall that the general structure of the argument for each principle is:

1. start with a general A-constraint: a constraint on the representeds representing an aspect of time-determination needed for unity
2. derive from the original A-constraint a further A-constraint, using the additional premise that time is not directly perceived, requiring the existence of a certain relation with certain properties
3. move to a B-constraint: a constraint on the representings, on the activity of combination
4. move to a C-constraint: a constraint on the rules used to produce the activity of combination

When translating these constraints into our logic-programming formalism, we treat the different stages differently. Constraints 1 and 3 are not encoded explicitly. If we satisfy 2, then we automatically satisfy 1. If we satisfy 4, then we automatically satisfy 3. We shall focus, therefore, on constraints 2 and 4. Constraint 2 is translated into a *relational* constraint on the set of stable models. Constraint 4 is turned into a requirement on the *structure* of the rules that are generated.

13.3.4 Translating Constraint 2 into a Relational Constraint

Each principle requires the existence of a certain relation satisfying certain properties. But the requirements on the various relations are *interdependent*: it is not possible to formulate the requirement on one relation without invoking some of the others. So, to specify the constraints in the First and Third[17] Principles, we will quantify over four binary relations on object-slices:

– PartOf(x, y): object-slice x is part of object-slice y (this implies they are simultaneous)
– SameObject(x, y): slices x and y are slices of the same object at different times

[17] I omit, for reasons of space, discussion of the Second Principle, the *Anticipations of Perception*. The Fourth Principle does not need its own relation.

– Succeeds(x, y): object-slice x is the predecessor slice of y
– Simultaneous(x, y): object-slices x and y are slices from the same moment in time

The entire set of constraints is one existential second-order sentence stating the existence of four binary relations that together satisfy the constraints, a sentence of the form:

$$\exists\ \mathsf{PartOf}, \exists\ \mathsf{SameObject}, \exists\ \mathsf{Succeeds}, \exists\ \mathsf{Simultaneous}...$$

13.3.4.1 Constraint 2 in the Axioms of Intuition

Constraint 1 of the *Axioms of Intuition* is the requirement that all our representations must be grouped into *moments*. This is the first aspect of time-determination: the collection of moments together constitute the entire time-series (Kant 1781)(B184, A145). The *Axioms of Intuition* are responsible for:

> the generation (synthesis) of time itself in the successive apprehension of an object (Kant 1781)(B184, A145)

Now, in Constraint 2, we move to a requirement that a certain sort of relation must exist: there must be a PartOf relation that allows the apprehensions to be grouped into moments.

The justification for this derived constraint is that, since moments cannot be directly perceived,[18] the only way a moment can be represented is as a concept derived from a relation. Define PartOf* as the transitive closure of PartOf. Constraint 2 requires that there is one largest element (the totality) for each moment such that every apprehension (object-slice) at that moment is PartOf* the totality element:

$$\forall x\ \exists! y\ \mathsf{Simultaneous}(x, y) \wedge \forall z\ \mathsf{Simultaneous}(z, y) \rightarrow \mathsf{PartOf}^*(z, y)$$

Now we can derive an equivalence relation InSameTotalityAs as the reflexive transitive closure of PartOf, and we can define moments as equivalence classes of this InSameTotalityAs relation: a **moment** is a maximal set of apprehensions (object-slices) S such that, for every x, y in S, InSameTotalityAs(x, y).

13.3.4.2 Constraint 2 in the Analogies

Common to all three of the *Analogies* is one high-level constraint: there must be a total ordering on moments in time (Kant 1781)(A145, B184-5).

[18] This is claimed explicitly in a marginal note to the first edition.

Now time is not directly perceived. Our sensations do not arrive with a convenient time-stamp we can inspect. In order to *construct* an ordering on moments in time, we must construct an ordering on object-slices. In fact, we need three relations between object-slices:

- SameObject
- Simultaneous
- Succeeds

The *First Analogy* focuses on the SameObject relation, an equivalence relation between object-slices. A pair of object-slices satisfy the SameObject relation if they are different slices of the same persisting space-time worm. This relation is connected to the relations of the other two *Analogies*:

- if $\mathsf{Succeeds}^*(x, y)$, then $\mathsf{SameObject}(x, y)$
- if $\mathsf{Simultaneous}(x, y)$ but x and y are distinct apprehensions, then it cannot be that $\mathsf{SameObject}(x, y)$

Constraint 2 for the *First Analogy* is the requirement that persisting objects (substances) exist at every moment in time:

> For every object x at moment m, for every other moment $m' \neq m$, there must exist a unique y at m' such that $\mathsf{SameObject}(x, y)$

But, since time cannot be directly perceived, we need to re-express this requirement without making use of explicit moments of time. One way to re-express the requirement is:

$$\forall x\ \forall y\ \neg\mathsf{Simultaneous}(x, y) \rightarrow \exists!z\ \mathsf{SameObject}(z, x) \wedge \mathsf{Simultaneous}(z, y)$$

The Second and Third *Analogies* impose constraints on the Succeeds strict ordering and the Simultaneous equivalence relation:

$$\forall x\ \forall y\ \mathsf{SameObject}(x, y) \rightarrow \mathsf{Simultaneous}(x, y) \vee$$

$$\mathsf{Succeeds}^*(x, y) \vee \mathsf{Succeeds}^*(y, x)$$

Here, $\mathsf{Succeeds}^*$ is the transitive closure of Succeeds.

The Second and Third *Analogies* also require a more general constraint on *any* two object slices x and y, even if they are not part of the same object. Define a derived relation $<$ on object-slices as the smallest transitive relation satisfying:

- if $\mathsf{Succeeds}(x, y)$ then $x < y$
- if $x < y$ and $\mathsf{Simultaneous}(x, x')$ and $\mathsf{Simultaneous}(y, y')$, then $x' < y'$

In other words, the $<$ relation relates object-slices that are not necessarily part of the same persisting object. Intuitively, it means x occurs at an earlier moment than y. Now the more general requirement is:

$$\forall x\ \forall y\ \mathsf{Simultaneous}(x, y) \vee x < y \vee y < x$$

13.3.5 Translating Constraint 4 into a Restriction on the Structure of the Rules

In this section, we translate Constraint 4 into a condition on the *structure* of the rules that are generated. These constraints are translated into a set of templates[19] that the logic programs are built from.

13.3.5.1 The Structure of Rules in the Axioms of Intuition

The types of rule we need to generate to satisfy the *Axioms of Intuition* are rules of composition, rules that allow us to create object-slices from other object-slices. The format of this rule is:

$$\forall x_1, \ldots, x_n \; \phi(x_1, .., x_n) \rightsquigarrow \exists y \bigwedge_{i=1}^{m} \psi_i(y, x_1, \ldots, x_n)$$

This rule allows us to construct a composite object-slice, y, satisfying multiple properties $\psi_1, \ldots, \psi_m$, if the slices $x_1, \ldots, x_n$ satisfy ϕ. Here, PartOf may be one of the ψ_i. Note that this is defeasible implication: this rule permits us to infer the consequent, but does not necessitate it. Sometimes, there will be many possible defeasible rules of composition, not all of which are compatible. For example, this group of sensory readings could be interpreted as representing a nose; or it could be interpreted as representing an ear; but both defeasible rules could not fire at once with the same substitutions for variables $x_1, \ldots, x_n$.

13.3.5.2 The Structure of Rules in the First Analogy

The *First Analogy* requires that there are rules of composition that produce object-slices for objects at moments at which they cannot be perceived. We need a rule such that, for each slice x and each moment m, it produces a slice z at moment m such that SameObject(z, x). This rule of the "productive imagination" (Kant 1781)(B154) has the form:

$$\forall x, y \; \phi(x, y) \rightsquigarrow \exists z \; \mathsf{SameObject}(z, x) \wedge \mathsf{Simultaneous}(z, y)$$

This is an instance of the general structure of rules of composition in Sect. 13.3.5.1 above.

[19] This is called a "language bias" in the program induction literature.

13.3.5.3 The Structure of Rules in the Second Analogy

To count the subjective sequence of apprehensions as objectively successive, we must apply a causal rule that explains why this particular object's fluent properties changed in the way that they did, and also necessitates other object's fluent properties changing in similar ways in similar situations.

Define a three-place relation $\mathsf{Causes}(x, y, p)$ relating two object-slices x and y and a reified proposition p. This causal relation determines both which propositions become true and the $\mathsf{Succeeds}$ relation between slices:

$$\mathsf{Causes}(x, y, p) \rightarrow \mathsf{Succeeds}(x, y)$$
$$\mathsf{Causes}(x, y, p) \rightarrow \mathsf{Is}(p)$$

Here, $\mathsf{Is}()$ is a predicate holding of reified propositions exactly when those propositions are true.

Now the template for causal relations is:

$$\forall x, y\ \phi(x, y) \rightarrow \mathsf{Causes}(x, y, \psi(x, y))$$

where $\psi(x, y)$ is a reified proposition: a term describing the propositional result of the causal interaction in which object-slice x succeeds to y. For example, if a melting candle gets smaller by 1 cm each time-step, then the causal rule would relate two successive apprehensions of the candle, x and y, and $\psi(x, y)$ would state the height of the candle-slice y as a function of the height of the candle-slice x.

13.3.5.4 The Structure of Rules in the Third Analogy

The only way we can count our apprehensions as simultaneous is if we construct an interaction rule that reciprocally determines these particular slices' fluent properties.

Define a three-place relation $\mathsf{Interacts}(x, y, p)$ relating two object-slices x and y and a reified proposition p. The $\mathsf{Interacts}$ relation determines both which propositions become true and the $\mathsf{Simultaneous}$ relation between slices:

$$\mathsf{Interacts}(x, y, p) \rightarrow \mathsf{Simultaneous}(x, y)$$
$$\mathsf{Interacts}(x, y, p) \rightarrow \mathsf{Is}(p)$$

Again, $\mathsf{Is}()$ is a predicate holding of reified propositions exactly when those propositions are true.

Now the template for interactions is:

$$\forall x, y\ \phi(x, y) \rightarrow \mathsf{Interacts}(x, y, \psi(x, y))$$

where $\psi(x, y)$ is a reified proposition: a term describing the result of the interaction between simultaneous object-slices x and y. For example, if a touch-sensitive sensor is turned on when a moving object presses against it, the two slices x and y are the slices of the sensor and the moving object, and the proposition $\psi(x, y)$ is the fact that the sensor is turned on.

13.3.5.5 The Structure of Rules for the Postulates of Empirical Thought

In order to rule out certain propositions as impossible, we must construct necessary rules of incompatibility and rules of entailment that, when applied, rule out certain configurations:

> The schema of possibility is the agreement of the synthesis of various representations with the conditions of time in general (e.g., since opposites cannot exist in one thing at the same time, they can only exist one after another). (Kant 1781)(A144, A184)

These necessary rules of connection have the form:

$$\forall x_1, \ldots, x_m \neg(\phi_1(x_{1_1}, \ldots, x_{1_k}) \wedge \ldots \wedge \phi_n(x_{n_1}, \ldots, x_{n_k}))$$

For example:

$$\forall x, y, \neg(\mathsf{Nose}(x) \wedge \mathsf{Ear}(y) \wedge \mathsf{SameObject}(x, y) \wedge \mathsf{Simultaneous}(x, y))$$

13.3.6 Searching for Non-monotonic Logic Programs that Satisfy the Relational and Structural Constraints

Now, given a set of sensory perturbations (represented as logical atoms), the Kantian cognitive agent finds a set of rules, according to the templates specified by the structural constraints (detailed in Sect. 13.3.5), that generate some stable model satisfying the relational constraints (detailed in Sect. 13.3.4). If the agent finds such a set of rules, it has constructed a coherent interpretation of its sensory given, making a unity out of the plurality of sensings. Perception, then, is a form of program synthesis[20]: finding a logic program that, when applied, makes sense of the sensory perturbations we are given.

The Kantian agent's task is related to, but different from, the standard Inductive Logic Programming (ILP) task. In ILP, the learning agent is given a background theory, B, and two sets of examples: a set P of positive examples of the target

[20]Contrast with Shanahan (2005), who sees perception as a form of *abduction*.

predicate, and a set N of negative examples of the target predicate. The task of ILP[21] is to produce a logic program, R, a set of rules, such that

- $B \cup R \models p$, for all $p \in P$
- $B \cup R \not\models n$, for all $n \in N$

The Kantian agent's task is rather different from the ILP problem because the Kantian agent is not given a set of positive and negative examples of a target predicate. The Kantian agent performs *unsupervised learning*: he is given some input, of course. But the input he is given is a set of atoms representing the states of his sensors. He is not given a supervised signal of what he is supposed to learn. *He is free to construct any set of rules he chooses*[22] – as long as the rules satisfy the structural constraints (Sect. 13.3.5) and the resulting rule-set has a stable model satisfying the relational constraints (Sect. 13.3.4).

In the current implementation, the machine searches through the space of logic programs using a variant of an ILP algorithm described by Corapi et al (2012). Each structural constraint described in Sect. 13.3.5 above is converted into a template for generating ASP clauses, using negation-as-failure to implement the defeasible rules. The search-space is explored through a form of iterative deepening. I omit the technical details for reasons of space.

13.4 Experiments

I took a simplified[23] version of the Kantian cognitive agent, described above, and tested it in two domains: a two-dimensional grid world, and a one-dimensional world of string sequences. These are preliminary, proof-of-concept experiments. But they do show that a Kantian cognitive agent is capable of making sense of sensory data without supervision, from a tiny handful of examples.

[21] The problem description for finding *non-monotonic* logic programs from positive and negative examples is actually somewhat more complicated, as there may be multiple models, each with their own positive and negative instances. See Law et al (2014) for details.

[22] Hence Kant's emphasis on *spontaneity*: the Kantian agent is both less free (because he can *only* perform actions by applying rules) and more free (because he can construct *any set of rules he likes*) than the empiricist can possibly imagine.

[23] There are two major simplifications in the current implementation. The first is that the spatial framework needed to satisfy the *Axioms of Intuition* is given in advance, pre-specified, hand-coded. The agent is told that he is operating in a 2-dimensional grid world. The second major simplification is that the constraints involved in the *Anticipations of Perception* are ignored altogether: in the initial implementation, time is modelled as a series of discrete points, rather than being dense. In future work, I plan to overcome these limitations.

Table 13.1 Sensory data

	Sensor w	Sensor x	Sensor y	Sensor z
t_1	Off	Off	Off	Off
t_2	On	Off	Off	Off
t_3	Off	On	Off	Off
t_4	Off	Off	On	Off
t_5	Off	Off	Off	On
t_6	Off	Off	Off	Off

13.4.1 Making Sense of the Grid World

Imagine a robot in a simple grid world. The robot has an array of four sensors, w, x, y, z, arranged in a line. The robot is given sensory data for six time steps, as shown in Table 13.1. Here, the sensors are turned on, one at a time, from left to right. To get a feeling for the sensory world of the robot, clench one fist, close your eyes and ask someone to drag some object (perhaps, a pen) along your knuckles. The sensations you will get are similar[24] to the sensations the robot receives.

The robot must construct a set of rules that, when applied, makes sense of this sensory data. The robot has to construct *some* explanation of the sensations it is given. Since the on/off predicates are *fluent* properties of the sensors, there has to be, according to the Second and Third *Analogies*, some rule to explain the changing values of the fluent. The robot has some work to do to construct an explanation of the fluent values changing.

When the Kantian machine is given this sequence of sensory data, the first explanation it finds is this: there is an object m moving from left to right in the row directly above the sensors. (see Fig. 13.3). Whenever the object is directly above a sensor, it makes the sensor turn on (indicated in Fig. 13.3 by a shaded block). Otherwise, the sensor remains off.

This explanation requires the following rules to be constructed:

- a rule of composition, according to the template for the *Axioms of Intuition* in Sect. 13.3.5.1, constructing an object-slice for m at position $(x, y - 1)$ for every sensor at (x, y) that is turned on: we count the sensor's being on as evidence for the existence of an object-slice at the position directly above the sensor
- a rule of composition, according to the template for the *First Analogy* in Sect. 13.3.5.2, generating an object-slice of the moving object for each time step when it is not perceived by the sensors (in this case, for time-steps 1 and 6)
- a causal rule, according to the template for the *Second Analogy* in Sect. 13.3.5.3, describing how the moving object moves one square from left to right each time step

[24] One important difference is that your tactile sensations are much more fine-grained: you receive a number of intermediate sensations as the object moves between your four knuckles. The robot just has four discrete boolean sensors (one for each knuckle).

Fig. 13.3 One interpretation of the sensory data. The sensors (w, x, y, z) are shown at each time step. E.g. $w1$ means sensor w at time step 1. The interpretation that the Kantian machine finds is that there is a moving object, m, moving from left to right, in the row above the sensors. The moving object turns on the sensor when the mover is directly above the sensor. Note that the moving object is not directly perceived; it is constructed (or invented) to make sense of the data

– an interaction rule, according to the template for the *Third Analogy* in Sect. 13.3.5.4, describing how the object presses on a sensor when it is directly above it, causing it to turn on

Note that, although these rules were constructed simply to make sense of the sensory data, they also allow us to infer that the moving object was at position (1, 1) at time t_1 (even though we cannot sense the object at that time) and to predict that the moving object will be at position (6, 1) at time t_6 (even though we cannot sense the object at that time). This is as it should be: a Kantian interpretation of the present will always also allow us to predict the future and retrodict the past.

To satisfy the *Axioms of Intuition*, the Kantian agent constructs the following defeasible rules of composition describing the moving object that is above the sensor that is turned on:

```
pred1(skf1(X)) :-
     is(on(X), T),
     is(at(X, P1), T),
     above(P2, P1),
     not -pred(skf1(X)).

is(at(skf1(X), P2), T) :-
     is(on(X), T),
     is(at(X, P1), T),
     above(P2, P1).
     not -is(at(skf1(X), P2), T).
```

These clauses generate an appearance from a sensor-reading: if we perceive that a sensor is on, then count its being on as evidence of the existence of another object, a moving object, that is above the sensor.

These rules are generated from the template in Sect. 13.3.5.1. The `skf1(.)` function is a skolem function replacing the existentially quantified variable y in that template.

Here, `pred1` is a gen-sym, a new predicate replacing ψ in the template of Sect. 13.3.5.1. The meaning of this predicate was not chosen in advance by an engineer – rather, it was constructed during the program-synthesis process. The meaning of this generated predicate is *entirely determined by its inferential role* in the clauses that were constructed by the machine. Examining all the rules in which it appears, it is clear that `pred1` is used to denote a new type of object: a type that is distinct from sensors. The various rules generated by the machine together define the behaviour of this new type of object: it exists in the row above the sensors, it moves from left to right, and when it is directly above a sensor, it turns it on.

To satisfy the *First Analogy*, the Kantian agent constructs the following clauses:

```
same_object(skf2(X,Y), X) :-
     appearance(X),
     appearance(Y),
     not exists_appearance_at_time(X, Y).
```

```
simultaneous(skf2(X,Y), Y) :-
    appearance(X),
    appearance(Y),
    not exists_appearance_at_time(X, Y).

pred1(skf2(X,Y)) :-
    appearance(X),
    pred1(X),
    not exists_appearance_at_time(X, Y).
```

These rules say: if we have already constructed an appearance X satisfying `pred1`, and we have no corresponding slice of X for the moment at which Y exists, then posit the existence of an unperceived slice, simultaneous with Y and conclude that it is also a `pred1`. These clauses are generated from the template in Sect. 13.3.5.2. Here, `skf2(.)` is a skolem function replacing the existentially quantified variable z in that template.

To satisfy the *Second Analogy*, the Kantian agent constructs the clause:

```
causes(X1, X2, at(X2, P2)) :-
    is(at(X1, P1)),
    left(P1, P2),
    pred1(X1),
    pred1(X2),
    not -succeeds(X1, X2).
```

This rule states that the moving object moves right by one grid square each time step. To satisfy the *Third Analogy*, it constructs:

```
interacts(X, Y, on(Y)) :-
    is(at(X, P1)),
    is(at(Y, P2)),
    above(P1, P2),
    pred1(X),
    sensor(Y),
    not -simultaneous(X, Y).
```

This rule states that, if a moving object is above a sensor, the moving object presses on the sensor, turning it on.

These rules, when applied to the sensory data, produce a stable model that satisfies the relational constraints in Sect. 13.3.4. The Kantian agent has found a coherent unified interpretation of its sensory data by generating instances of the a priori relations (PartOf, SameObject, Succeeds, Simultaneous) that connect the plurality of sensory perturbations into one cohesive unity. It is worth stressing that the Kantian agent has found a satisfying unified interpretation despite only being given a tiny amount of experiences: it had to construct a unified interpretation given a sequence of only six time-steps.

13.4.2 Making Sense of a One-Dimensional String of Letters: A Verbal-Reasoning Task

Hofstadter introduced the **Seek Whence** problem-set in Hofstadter (2008). In this task, the player is given a sequence of symbols, for example:

a, a, b, b, c, c, d, d, ...

The player needs to guess the next symbol. The *only* knowledge the player can use is the successor[25] relation: a $\prec$ b $\prec$ c $\prec$ ... Here are some example problems:

- a, b, c, d , *e*
- a, a, b, b, c, c, d, d , *e, e*
- a, k, b, k, k, c, k, k, k, d, k , *k, k, k*
- b, a, b, b, b, b, b, c, b, b, d, b, b, e, b , *b, f, b*

The Kantian agent, when confronted with a Seek Whence letter sequence, is forced (by the *Axioms of Intuition*) to interpret its sensory data in terms of moments of time, composed of apprehensions (object-slices) that are spatially related to each other. It is forced (by the *First Analogy*) to connect the apprehensions together into enduring objects, persisting through time. It is forced (by the *Second Analogy*) to interpret changes to these enduring objects in terms of causal rules that explain those changes.

The Kantian agent, in other words, is *doomed* to reinterpret its sensory data in terms of objects persisting through time, changing state according to intelligible causal laws. Surprisingly, the Kantian constraints are enough, on their own, to achieve human-level success on these verbal-reasoning tasks. Consider the following example:

b, a, b, b, b, b, b, c, b, b, d, b, b, e, b , *b, f, b*

Hofstadter called this example the "theme song" of the Seek Whence project, because of its ambiguity. The long string of *b*'s encourages us to mis-parse the sequence, while the true parsing Hofstadter intends is a sequence of triples, *bxb* for increasing x.

Figure 13.4 shows the results of the Kantian machine's deliberations on this particularly tricky Seek Whence problem. The Kantian agent parses this sequence in the way Hofstadter intends. Note how the sequence is interpreted as three objects, persisting over time, changing state in different ways. The left and right objects (Objects 1 and 3 in the diagram) remain the same, while the middle object (Object 2) increases its value every time-step. Note the causal rule (`update`) that is constructed to explain the state change of the middle object. Making sense of the letter sequence means constructing a program that, when applied, reinterprets the sequence as a set of objects, persisting through time, changing state according to intelligible laws.

[25] We assume, for simplicity, that the alphabet is cyclic, so that the successor of z is a.

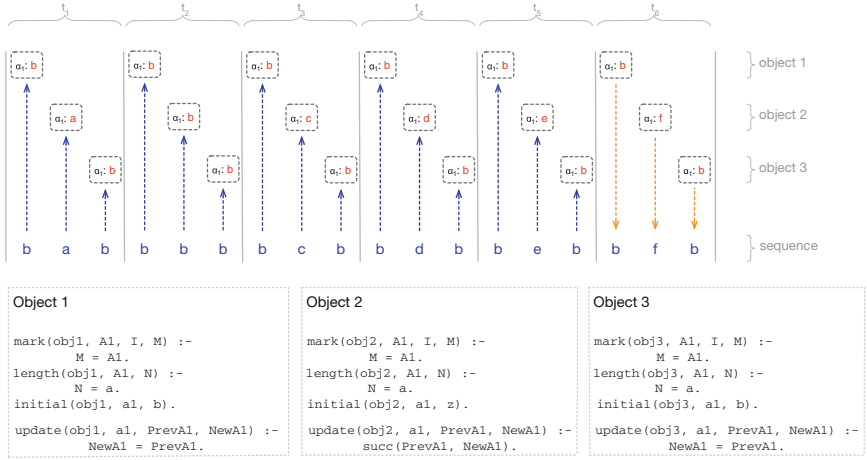

Fig. 13.4 Applying the Kantian machine to a tricky seek whence problem

The Kantian agent was tested on three data-sets: the "Blackburn Dozen" from Meredith (1986), the "Hofstadter Fifteen" from Hofstadter (2008), and the C-test from Hernandez-Orallo and Minaya-Collado (1998). Overall, the Kantian agent scored 86% correct[26] on the three data-sets, reaching human-level performance. This compares favourably with the only other known attempt to solve this problem, in Meredith (1986), that achieved 25% (Fig. 13.5).

13.5 Conclusion

The modern debate between deep-learning practitioners and advocates of logic-based approaches resembles the eighteenth century debate between empiricists and rationalists. One of Kant's driving forces was reconciliatory: to capture the insights of both empiricism and rationalism in one unified system. It seems at least possible that his work, his extraordinarily ambitious system, might have something useful to say about the modern-day reincarnation of this old debate.

This paper represents a first, tentative step towards a full computer implementation of Kant's vision: a self-legislating agent, bound by the rules he has

[26] We need to be careful with notions of "correctness" in sequence induction tasks. There are always infinitely many ways of continuing a finite series, even if some appear more "natural" to us than others. In the case of the "Blackburn Dozen" and the "Hofstadter Fifteen", the authors specified the intended continuation. I did not use these intended continuations when evaluating correctness. Instead, I gave the questions to 100 people, as an online form, and took the mode as the "correct" continuation. The Kantian constraints provide a way of formally specifying what is "natural" about the "natural" continuations.

Sequence	Human	Kantian Agent
b,b,b,c,c,b,b,b,c,c,b,b,b,c,c,...	b	b
b,a,a,b,b,b,a,a,a,a,b,b,b,b,b,...	a	a
b,a,b,e,b,a,a,a,e,b,b,b,b,e,a,...	e	-
b,c,a,c,a,c,b,d,b,d,b,c,a,c,a,...	e	a
a,b,b,c,c,d,d,e,e,f,f,g,g,...	h	h
a,a,b,a,b,c,a,b,c,d,a,b,c,d,e,...	a	a
b,a,c,a,b,d,a,b,c,e,a,b,c,d,f,...	a	a
a,b,a,c,b,a,d,c,b,a,e,d,c,b,...	g	g
c,b,a,b,c,b,a,b,c,b,a,b,c,...	b	b
a,a,a,b,b,c,e,f,f,g,g,g,h,h,i,...	s	-
a,a,b,a,a,b,c,b,a,a,b,c,d,c,...	a	a
a,a,b,c,a,b,b,c,a,b,c,c,a,a,...	a	a
a,a,b,c,a,b,b,c,a,b,c,c,a,b,...	a	a
a,b,b,c,c,a,a,b,c,c,a,a,b,b,...	a	a
a,b,b,c,c,a,b,b,c,a,b,c,c,a,...	a	b

Fig. 13.5 The Kantian agent's performance on Hofstadter's dataset (Hofstadter 2008)

himself constructed, reinterpreting his sensory perturbations as a coherent unified experience of an external world.

The implemented system is able to perform *unsupervised learning*: making sense of its sensory input without labeled data or rewards. This system takes the raw sensory stimuli and creates a program that, when applied, constructs a coherent interpretation of its sensory world.

The system is able to learn *data-efficiently*, from a tiny amount of data, because of the strong prior knowledge built into the system. This prior knowledge is a set of general, domain-independent constraints on the types of rule that can be constructed. These constraints are taken directly from Kant's *Analytic of Principles*: constraints that must be satisfied by any agent who seeks to reinterpret his sensory perturbations as a coherent whole, unified in the medium of time.

References

Chalmers, D.J., R.M. French, and D.R. Hofstadter. 1992. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence* 4(3): 185–211.

Corapi, D., A. Russo, and E. Lupu. 2010. Inductive logic programming as abductive search. In: ICLP (Technical Communications), 54–63.

Corapi, D., A. Russo, and E. Lupu. 2012. Inductive logic programming in answer set programming. In: *Inductive Logic Programming*, 91–97. Heidelberg/New York: Springer.

Frege, G., P. Geach, and M. Black. 1980. 'Über Sinn und Bedeutung', in Zeitschrift für Philosophie und philosophische Kritik, Translated as 'On Sense and Reference' by M. Black in Translations from the Philosophical Writings, 100: 25–50. Oxford: Blackwell, third edition.

Gelfond, M., and V. Lifschitz. 1988. International Conference on Logic Programming. The stable model semantics for logic programming. In: *ICLP/SLP*, vol. 88, 1070–1080.

Goodman, N.D., J.B. Tenenbaum, J. Feldman, and T.L. Griffiths. 2008. A rational analysis of rule-based concept learning. *Cognitive Science* 32(1): 108–154.

Graves, A., et al. 2012. Supervised sequence labelling with recurrent neural networks, vol. 385. University of Toronto, Springer.

Haugeland, J. 1990. The intentionality all-stars. *Philosophical Perspectives* 4: 383–427.

Hernandez-Orallo, J., and N. Minaya-Collado. 1998. Engineering of Intelligent Systems, A formal definition of intelligence based on an intensional variant of algorithmic complexity. In: *Proceedings of International Symposium of Engineering of Intelligent Systems (EIS98)*, February 11–13, 146–163.

Hofstadter, D.R. 2008. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. New York: Basic Books.

Hofstadter, D.R., M. Mitchell, et al. 1994. The copycat project: A model of mental fluidity and analogy-making. *Advances in Connectionist and Neural Computation Theory* 2(31–112): 29–30.

Hutter, M. 2007. On universal prediction and Bayesian confirmation. *Theoretical Computer Science* 384(1): 33–48.

Jordan, C., and L. Kaiser. 2013. Learning programs as logical queries. In: *The ICALP 2013 Satellite Workshop on Learning Theory and Complexity*, (ICALP is the "International Colloquium on Automata, Languages and Programming").

Kant, I. 1781. *Critique of Pure Reason*, Trans. P Guyer. Cambridge University Press.

Kowalski, R., and M. Sergot. 1989. A logic-based calculus of events. In: *Foundations of Knowledge Base Management*, 23–55. Berlin: Springer.

Lake, B.M., R. Salakhutdinov, and J.B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266): 1332–1338.

Law, M., A. Russo, and K. Broda. 2014. Inductive learning of answer set programs. In: *Logics in Artificial Intelligence*, 311–325. Cham: Springer.

Longuenesse, B. 1998. *Kant and the Capacity to Judge*. Princeton: Princeton University Press.

McCarthy, J. 1963. Situations, actions, and causal laws. Technical Report, DTIC Document.

Meredith, M.J.E. 1986. Seek-whence: A model of pattern perception. Technical Report, Indiana University, Bloomington (USA).

Mitchell, M. 1993. Analogy-making as perception: A computer model. Cambridge: MIT Press.

Muggleton, S.H., D. Lin, and A. Tamaddoni-Nezhad. 2015. Meta-interpretive learning of higher-order dyadic datalog: Predicate invention revisited. *Machine Learning* 100(1): 49–73.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13(1): 81–132.

Sellars, W. 1968. Science and metaphysics: Variations on Kantian themes. Ridgeview Publishing Company, Springer.

Shanahan, M. 2005. Perception as abduction: Turning sensor data into meaningful representation. *Cognitive Science* 29(1): 103–134.

Sloman, A. 2008. Kantian philosophy of mathematics and young robots. In: *International Conference on Intelligent Computer Mathematics*, 558–573. Berlin/Heidelberg: Springer.

Tenenbaum, J.B. 2000. Rules and similarity in concept learning. *Advances in Neural Information Processing Systems* 12: 59–65.

Waxman, W. 2013. *Kant's Anatomy of the Intelligent Mind*. Oxford: Oxford University Press.

Wittgenstein, L. 1958. *The Blue and Brown Books*. Oxford: Blackwell.

Part V
Moral Dimensions of Human-Machine Interaction

Chapter 14
Machine Learning and Irresponsible Inference: Morally Assessing the Training Data for Image Recognition Systems

Owen C. King

Abstract Just as humans can draw conclusions responsibly or irresponsibly, so too can computers. Machine learning systems that have been trained on data sets that include irresponsible judgments are likely to yield irresponsible predictions as outputs. In this paper I focus on a particular kind of inference a computer system might make: identification of the intentions with which a person acted on the basis of photographic evidence. Such inferences are liable to be morally objectionable, because of a way in which they are presumptuous. After elaborating this moral concern, I explore the possibility that carefully procuring the training data for image recognition systems could ensure that the systems avoid the problem. The lesson of this paper extends beyond just the particular case of image recognition systems and the challenge of responsibly identifying a person's intentions. Reflection on this particular case demonstrates the importance (as well as the difficulty) of evaluating machine learning systems and their training data from the standpoint of moral considerations that are not encompassed by ordinary assessments of predictive accuracy.

Keywords Machine learning algorithms · Image recognition systems · Training data · Responsible AI judgment · Ingrained responsibility · Modular responsibility · Intention ascription

O. C. King (✉)
Department of Philosophy, University of Twente, Twente, The Netherlands
e-mail: o.c.king@utwente.nl

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_14

14.1 Introduction: Humans and Computers Drawing Conclusions Responsibly

Consider Ned, who does not know the difference between a peanut and a cashew. In fact, *cocktail nut* is the most specific category in this region of Ned's gastronomic conceptual taxonomy. Now suppose I've taken it upon myself to teach Ned to see the difference. So, I show him five labeled photos of peanuts and five labeled photos of cashews. Then I show him a new picture of a nut without a label. He confidently says, "Peanut!" and he is correct. I show him a bunch more new photos, and he identifies all the peanuts and cashews correctly. Mission accomplished. Ned has learned to visually discriminate peanuts and cashews.

Machine learning systems for image recognition operate much the same way. They are fed sets of images paired with descriptions, which are the *training data*. And then the systems generate descriptions for (or match pre-given descriptions to) new images. It amounts to an advance in image recognition when a system can draw more accurate conclusions than previous systems on the basis of the same training data. But this is not the only sort of improvement possible. Training data can be improved, too. The set of images could include more relevant variety, or the descriptions could be more accurate, or the data set could just be more voluminous. In our example of Ned, better training data might mean teaching him using sharper images of peanuts and cashews. Probably images that showed differences in the textures of the two types of nuts, all else equal, would be more helpful to him than images that lacked this level of detail.

It is tempting to think that if one set of training data yields computer systems that draw more accurate conclusions than those from systems trained on other data, then the data set that yields the more accurate systems is better. But I do not think this is the whole story. As machine learning systems, such as image recognition systems, become more and more sophisticated with wider and wider application, it is not just the accuracy of the conclusions that matters. Just as a judgment pronounced by a human might have been irresponsible, despite its accuracy, computer systems also can draw conclusions irresponsibly though accurately. And this irresponsibility can be due to the data on which the systems were trained.

Here are a couple cases of human judgment that exemplify the kind of worry I have in mind. Suppose we have an image of a man running behind a running woman who has a frightened look on her face. Suppose I look at the image and say, "He's trying to hurt her!" Well, I might very well be correct. But, clearly, my judgment has overshot my evidence. What if the man and the woman are both fleeing from some other menace? Or suppose we have a photo of a man and a woman, both finely dressed, smiling as they sit at a candle-lit table with an elegant dinner laid out before then. If I say, "They're on a date," then my judgment has gone too far again. Perhaps they're just friends; it might even be that they're both gay.

Note that the worry here is not just epistemic. We might even suppose that images that look relevantly like the first one 98% of the time really do picture one person trying to hurt another, and we might suppose that 98% of images relevantly like

the second really do depict dates. The irresponsibility involved here is more about *failure of respect* than about a lack of evidence; it is *more ethical than epistemic*.

Suppose I am barely acquainted with the two people—call them Jack and Cleo—shown in the dinner picture. And suppose I was at that restaurant and happened upon that very scene. If I ran into Jack at the coat check, I wouldn't say, "How's the date going?" Expressing the judgment that they're on a date would be presumptuous. And my embarrassment will be fitting if Jack says, "I'm gay, you idiot." But even if Jack and Cleo are indeed on a date, my comment would be no less presumptuous. The problem is that my inference was based on a superficial pattern they seemed to fit, rather than on any intentions they had expressed or any other facts about them as individuals.

Now if we have an image recognition system trained on data that include judgments like those in the two examples I just described—the example of the running people and the dinner example—then the irresponsible (though perhaps quite accurate) judgments will affect the way the system operates subsequently. Irresponsible judgments in the training data are likely to yield irresponsible conclusions at runtime.

The main issue here extends beyond just image recognition. The general issue is about the responsible use of AI systems capable of making judgments that might carry some moral weight. How ought the developers of these systems ensure that the systems judge responsibly? One option is to train the systems just for statistical accuracy, and then add an extra layer of processing to ensure that the judgments are applied responsibly. A second approach is to train the responsibility into the system from the beginning, by ensuring that the set of training data does not encode some pattern of irresponsibility. We can think of these approaches as *modular responsibility* and *ingrained responsibility*, respectively.

In the rest of this paper I will consider how we might achieve ingrained responsibility for machine learning systems, especially image recognition systems, that draw conclusions about what actions persons perform. This focus is attractive because of the present and ongoing advances in the development of such systems. In general, I suspect that it is prudent for us to prefer ingrained responsibility over modularization. But, as we will see, the temptation to modularize responsibility will be strong.

14.2 Presumptuous Judgment

Before returning to issues about image recognition and machine learning, it is worth elaborating the central moral concern here. The basic worry is that some judgments about a person's actions may be objectionably presumptuous. My goal is not to give a comprehensive account of presumptuousness or the reasons it is objectionable, but I hope to say enough about it to illuminate the sort of worry I have in mind.

We can say a judgment of a person's intentions is *presumptuous* when the intentions were ascribed on the basis of superficial features of the person, instead

of on the basis of the person's own individual profile of past and present mental states. This way of characterizing presumptuousness is not intended to be a precise definition that draws a sharp boundary around all the cases of presumptuous judgment. It is quite possible that our thinking about these issues is too hazy and mutable to make drawing a sharp boundary desirable or even feasible. Instead, what this characterization does is locate and orient presumptuous judgment with respect to types of possible evidential bases. The more a judgment of a person's intentions is based on facts about that particular individual's thoughts and desires—as manifested in, say, prior action or speech—the less presumptuous it is. The more the judgment is based on other characteristics of the person—especially general, population-wide patterns she seems to fit—the more presumptuous it is. I will not provide here a thorough defense of the claim that presumptuousness is morally problematic, but I will try to say a little bit to make the claim plausible.

First, it is worth observing that many among us (including myself) tend to be offended when people make unwarranted assumptions about our desires, goals, and intentions. Consider this scenario: Suppose I have an acquaintance, Silas, who is a bit overweight. I overhear a conversation in which Silas mentions that he has planned a trip to the beach several months from now. So I infer that Silas intends to lose weight. (He wouldn't want to look fat in his swimsuit, right?) Then, when a mutual friend is preparing for a dinner party, to which Silas and I are both invited, I suggest that she include only light fare on the menu, since (I believe) Silas is trying to lose weight. Now, as it turns out, Silas is not at all concerned about his weight. It would be fitting for Silas to be offended, or at least annoyed, at my presumptuousness. Note that the problem is not that my inference was terribly faulty from a purely epistemic standpoint; it was that I made an inference (which I then acted upon) about Silas's intentions, even though I did not know enough about Silas to do so responsibly. So, I should have withheld judgment, or at least abstained from acting on my judgment.

For another example, consider another scenario involving Silas. Suppose Silas decides to send his daughter to the local public high school instead of the nearby, expensive, private high school. Upon hearing about this, Silas's neighbor Albert infers that Silas is trying to save money. As it turns out, Silas's choice was motivated by his hope that his daughter will benefit from an education among a more inclusive group of students. Here again, it would be fitting for Silas to react with offense or annoyance at the presumptuous judgment.

Second, note the close link between presumptuousness and stereotypes. We can understand many stereotypes as constituted by shared patterns of presumptuousness. For example, imagine that Ravi is an Indian-American college student whose parents immigrated to America shortly before he was born. At college Ravi chooses pre-med as his major. Peter, Ravi's roommate, assumes that Ravi is aiming to become a doctor because of pressure from his demanding parents. It turns out that Ravi has always been interested in human biology and the practical applications of it. Peter's presumptuous judgment about Ravi's goals was a manifestation of a general stereotype Peter has accepted about Indian-Americans. Note that even if Peter had been correct about Ravi's motivations, basing his judgment on a stereotype

about Indian parents still would have been inappropriate. It is not hard to think of cases of stereotyping that are much more pernicious than this one.[1]

Finally, consider this not-too-far-fetched example, which is a bit more like the image recognition cases that are our main concern. Imagine Tara, who is an academic advisor at a large state university in the U.S. One of her duties is to have one-on-one meetings with incoming students to help them choose and register for courses during their first semester at college. Now, after a couple of years of conducting these meetings, Tara realizes that the meetings would be much more productive if she proposed a default schedule to the student at the beginning of each meeting. So, she tries this, and each student starts with a default schedule that includes Calculus I, First-year Writing and Composition, Problems of Philosophy, and Intro to the Life Sciences. At first, she does not customize the schedule for each student because she has little information on which to base any recommendations. Because of a poorly conceived information and record system at the university, she has just a photo of the student and the student's home address. However, after the first year of using her new system, and despite her dearth of background information, Tara happens to notice one regularity: Male students who hail from the northern part of the state and who are pictured in preppy attire always want to sign up for Intro to Business. So, Tara adjusts her system. For most of her advisees, she continues to offer that original default schedule. However, for her preppy, northern males, she includes the introductory business course in place of the life sciences course. After this adjustment, Tara's own personal records indicate that she has reduced her average meeting duration by 5%. So she makes the adjustment permanent.

Despite the increased efficiency from Tara's newly adjusted policy, it may strike us as suspect. But if there is a problem here, it is not inaccuracy or lack of evidence. The policy was devised on the basis of plenty of data, and it is even backed by some empirical confirmation. The problem is that she is predicting individuals' preferences (and using these predictions in ways that might influence them) on the

[1] A few clarifications about the relationship between stereotypes and presumptuous judgment may be helpful. First, not all cases of presumptuous judgment involve stereotypes. Stereotypes involve associating an individual with a group (Blum 2004; Beeghly 2015). But it is possible to make a presumptuous judgment without relying on a group association. For instance, I might make a presumptuous judgment about a person's intentions just on the basis of the assumption that her goals are the same as my own. Second, not all uses of stereotypes involve presumptuous judgments. This is simply because not all stereotypes are about persons' intentions. Finally, regarding the moral features of stereotypes and presumptuous judgments: Presumptuousness, all else equal, tends to be morally undesirable, but it's controversial whether this is true of all stereotypes. Beeghly (2015) argues that not all stereotyping is morally objectionable, and Lippmann (1922) saw positive and negative aspects of stereotyping. In contrast, Blum (2004) holds that stereotyping is always morally objectionable to some degree. My contention here, that presumptuous judgments manifest inadequate respect for persons as individuals, is consistent with Beeghly's explanation of when and how stereotypes fail to respect persons as individuals. However, my thinking about why such a failure of respect is morally objectionable shares more with Blum's analysis than with Beeghly's. In the context of the present paper—with its focus on the moral evaluation of training data for machine learning systems—it is enough for my purposes if at least some judgments are morally objectionable precisely because of their presumptuousness.

basis of the persons' conformity to a superficial pattern, and thus failing to treat them as individuals. The problem is a moral one.

If indeed presumptuousness of the sort I've been gesturing at is undesirable, we will not want our computer systems to issue presumptuous judgments. As already noted, one approach, the modular approach, would have us outfit our computer systems with an additional stage of processing which took the set of statistically founded judgments and filtered out the presumptuous ones. The ingrained approach, which I'm exploring here, would effectively apply a filter on the opposite end, removing presumptuousness from the training data. To see how this would work in the case of image recognition systems, we need to look a little more closely at these systems and how they are trained.

14.3 Image Recognition and Sources of Training Data

There are various kinds of image recognition tasks we may wish to have a computer perform. Given a photograph, we may wish to have a computer classify *what kind of scene* it is (for example, a desert or a grocery store) or identify *what objects* are pictured (for example, a camel or a cantaloupe). We might also wish to have the computer draw more nuanced conclusions—specifically about the relations among various elements and *what is happening* in the photograph (Fei-Fei and Li 2010). For example, we might like the computer to tell us that a camel is drinking from a spring or that a boy is adding a cantaloupe to his shopping cart.

Advances in computer vision in the last decade have begun to make automated scene classification and object identification more practical. And recently, new research has made headway in the third sort of task. Some new image recognition systems can tell, with some accuracy, how the objects in an image are related—reporting not just the *what*, but also the *what's going on*. This progress is the result of combining two branches of AI research: computer vision and natural language processing. The new image recognition systems integrate visual meaning and linguistic meaning in the same models, facilitating greater precision and subtlety in associating descriptions with images (Karpathy and Fei-Fei 2014; Vinyals et al. 2014).

At a basic level, recent innovations notwithstanding, the new AI systems operate on the same principles as their predecessors. The first step is usually to feed the systems large sets of data. It is from this training data that a system "learns" (i.e., creates a rich model of the data). It is only once some learning has taken place that the machine learning system becomes useful. (Whether the learning process continues once the system is in operation depends on the specific system and its implementation.) In the case of image recognition, the training data includes scores of images paired with descriptions. Different data sets include different images, and the form of the descriptions may vary as well—from single-word descriptions to multi-sentence paragraphs.

What are the sources of training data for image recognition systems? It is tempting to think we have an embarrassment of riches. The Internet, from professional media outlets to social media, provides a never-ending stream of captioned images. It is Big Data *par excellence*. Consider how e-commerce websites like Amazon and eBay analyze their unceasing streams of consumer behavior data in order to train their systems to make more intelligent product recommendations. Similarly, to train our image recognition systems, one might think that we just need to point them at the streams of captioned photos that perpetually pour from the likes of Facebook, Twitter, Flickr, Instagram, Pinterest, Imgur, etc.

But a bit of reflection shows that this approach is a non-starter. After all, why do people caption images in the first place? The goal is certainly not to give plain and literal, yet comprehensive, descriptions of the contents of the photos. Instead the goal is to tell us about the things not pictured—like important background information—that make the photo interesting. If a photo shows a chemist in her lab, the caption is likely to say who she is and what she studies. It will *not* say anything like this: "A woman with goggles and a white coat lifts a glass vessel containing blue liquid." Such a caption would be useless to us; we can notice all this (and much more) from a quick glance at the photo.[2] But this is exactly the kind of caption we need paired with our image if it is to be part of our training data. The point, then, is that the training data we need for image recognition systems—unlike paradigmatic big data applications in which the relevant data sets continually accrete through the everyday course of events—must be artificially created and collected.

Artificial creation of training data is a daunting task, but it's not quite as difficult as it might initially seem. Researchers and developers can simply hire people to describe photos. And with *crowdwork services*—like Amazon's Mechanical Turk—which crowdsource the completion of large sets of microtasks, it is fast and inexpensive to create large sets of training data. Researchers can define tasks and advertise them within Mechanical Turk, and then human workers (the "Turkers") find them and complete them. In 2009, computer vision researchers at the University of Illinois used Mechanical Turk to acquire human-generated descriptions for over 8000 images from the Flickr photo sharing website, in less than 12 days and at a cost of less than $1000 (Rashtchian et al. 2010). The result was a data set known as Flickr 8k, which includes approximately 8000 images paired with the descriptions written by Turkers (Hodosh et al. 2013). Thus, crowdwork takes care of the major practical obstacle in the way of training image recognition systems.[3] So, now we can begin worrying about ingrained responsibility—what it takes to make sure that none of the image labels in our training data express presumptuous judgments.

[2] As Hodosh et al. (2013) point out, "Gricean maxims of relevance and quantity entail that image captions that are written for people usually provide precisely the kind of information that could not be obtained from the image itself, and thus tend to bear only a tenuous relation to what is actually depicted."

[3] Though crowdwork raises ethical issues of its own (Marvit 2014).

14.4 Integrated Responsibility for Still Photographic Training Data

How could a group of workers—individuals paid to label images—produce training data that encodes responsible judgments about what people depicted in the pictures are doing? The simple answer is that the workers must adhere to strict instructions about the kind of descriptions they are to provide. If I am right that presumptuous judgments are morally objectionable, then the instructions should rule out presumptuous judgments. So, one option would be simply to instruct the workers to avoid presumptuousness.

But this sort of instruction is awfully abstract and not the most straightforward to operationalize on a case-by-case basis. Clearer instructions are required. As it turns out, Hodosh et al., the team that created the Flickr 8k data set, did an admirable job with their instructions. In a qualification test for workers who might write image descriptions, the researchers gave prospective workers this characterization of a good description:

> A good description...
>
> ...should provide an explicit description of prominent entities in the image.
> ...should not make unfounded assumptions about what is occurring in the image.
> ...should only talk about entities that appear in the image.[4]

The third and especially the second of these three clauses should serve to rule out many cases of presumptuousness. After all, part of what constitutes presumptuousness, as I've characterized it, is an inappropriately grounded judgment about what is motivating a person. So, my complaint is only that these instructions are not strict enough in what they prohibit. As we've seen, a judgment may be well-founded, in that it is statistically well-supported, yet presumptuous nonetheless. If presumptuousness is indeed undesirable, the rules for making assumptions about persons' actions should be more strict than the rules for making assumptions about other sorts of occurrences.[5] For example, the graphical data in an image depicting the view from a window looking out into a rainy day may be consistent with the unlikely possibility that the falling drops of water are coming from a sprinkler somewhere off to the side, but that wouldn't make the judgment that it's raining inappropriate. In order for our image recognition systems to be as useful as possible, we would prefer an image that appears to depict rain be described as depicting rain.[6] A 2% chance that it is not actually raining is not enough to withhold the

[4]This comes from the online appendix to Hodosh et al. (2013).

[5]This suggests another way to explain what is wrong with presumptuous judgment. To judge a person's mental states according to a standard like we would use for any other sort of judgment not involving persons, is to take what Peter Strawson (1962) called the "objective attitude" rather than the "participant attitude" toward the person.

[6]Of course, the image recognition system could report the falling water, and we could rely on some other process to infer from the falling water that it must be raining. But this would be to limit

Fig. 14.1 An image from the Flick 8k data set

judgment that it is raining. However, a 2% chance of error is enough to withhold the judgment that the dining man and woman are on a date.[7] That is because there is more to avoiding presumptuousness than making judgments with sufficiently high probability.

To instruct workers in such a way that their descriptions avoid presumptuousness, I propose the following addition to the instructions used by Hodosh et al.: *Do not give a description of an action such that the person could plausibly deny that that's what she was doing.* As with the original instructions, some vagueness remains. However, the meaning of the instructions can be demonstrated with examples. (And such examples could be included with the instructions to the workers.)

Consider Fig. 14.1, in which a woman and a young boy stand next to a table covered with various foodstuffs. This image, along with five English descriptions written by Turkers, is included in the Flickr 8k data set.

---

too much the capacities of image recognition systems. A scene can be one that *looks rainy*, and looking rainy may be both more intuitive and more useful information than the report that *it looks like water is falling from above*.

[7] There's nothing special about the specific probability values of 0.02 and 0.98, besides the former being small and the latter being large. These values are just convenient for purposes of illustration. Values of 0.01 and 0.99 or 0.05 and 0.95 would have worked just as well (although values that were too extreme or too moderate would indeed alter the examples).

The Turkers' descriptions of Fig. 14.1 were as follows:

1. A woman and a boy are making hamburgers in the kitchen.
2. A woman in a white shirt prepares a large meal of hamburgers.
3. A woman is holding a jar of mustard and a boy is looking at a tray of hamburgers.
4. The woman has a blue shirt on with a kid to her side, and she is making hamburgers.
5. Woman and young boy stand in a kitchen with a spread of burgers in front of them.

Among these five descriptions, only (3) and (5) would be acceptable according to the additional instruction I am proposing.[8] The others make presumptuous inferences about the woman's intentions. It is clear that the woman is holding a mustard jar and sticking some kind of utensil in another jar on the table. However, it is unclear what she intends to be doing. She might be just taking a hamburger for herself; or perhaps she is just sampling the mustard. (Returning to the point I noted earlier about the relationship between presumptuousness and stereotypes, it is worth wondering whether the Turkers would have written different descriptions if the picture had included an old man in a suit instead of a young woman in a casual blouse!)

Figure 14.2 is another image from the Flickr 8k data set.

The descriptions of Fig. 14.2 were as follows:

1. Four people are lining up to purchase tickets at the theater.
2. Four people standing outside of an outdoor ticket booth.
3. Four people wait outside in a line for ticket.
4. The man and woman at the window are turned around to the man and woman behind them.
5. Two men and two women standing at the window of a ticket booth.

Among these descriptions, (1) and (3) would be prohibited by the rule I have proposed. The reason is that they attribute intentions to the persons depicted. We are not in a position to know that these people are indeed trying to acquire tickets. They might be there just to ask a question, or perhaps they are in line to get a refund, not make a purchase at all.

Despite these examples of how the instruction I've proposed would have affected this data set, I must point out that the change would be very minor. If the workers writing descriptions had adhered to my proposed instruction, the Flickr 8k data set would *not* be very different than it is. That is because the workers attributed intentions to the individuals pictured fairly seldom. This is good news. It means that the data set is useful for training image recognition systems, without much risk of generating presumptuousness.

But now we are in a position to observe that this success comes with a cost. The fairly strict limit on what is allowed in the descriptions limits the scope of

[8] I do not intend this as a criticism of the Flickr 8k data set. Violations of the instruction I am recommending seem to appear only rarely in the data set. However, this image and the next are valuable for illustrating the worry I that is my focus.

Fig. 14.2 Another image from the Flick 8k data set

the judgments that can be produced by a system trained on such a data set. The descriptions of the activities in the training images are to be limited to, at most, the *overt behavior* of the persons pictured. So, the captions can describe *intentional actions* in only a very thin sense. For instance, we might say of a photo that it shows a woman kicking a soccer ball, but we cannot say that she is passing or shooting—at least not on the basis of a single still image. Necessarily missing is any attribution of aims, attempts, plans, or processes. And if the training data lacks these sorts of attributions, then a system trained on these data cannot possibly attribute them either. If we want a machine learning system to provide rich, informative descriptions of intentional actions, but also do so in a non-presumptuous way, then we will have to broaden the training data.

14.5 Theoretical Grounds for Ascribing Intentions?

We have seen that if we adhere strictly to the sort of principle I've advanced, descriptions generated by a system trained on data like the Flickr 8k data set will be limited in their informativeness. Such a system can offer very little in the way of responsible judgments about persons' intentional actions. However, many applications—indeed any applications designed to intelligently assist a user with the achievement of her goals—will need information about the user's intentions.

It is tempting here to fall back on a general idea about the basic conditions of successfully interpreting—making sense of the thoughts, behavior, and speech of—one another. Let me explain. W. V. Quine famously argued that radical translation—the process of translating the previously unknown language of a foreign speaker into one's own language—requires applying a principle of *rational accommodation*, what's more commonly known as a *principle of charity* (Quine 1960). The principle is required when trying to make headway in a situation in which the only evidence available to an interpreter is the overt behavior (including utterances) of the foreign speaker whose language the interpreter is trying to understand. The behavioral evidence will necessarily be compatible with many different translations, given the many different background beliefs the foreign speaker may hold. In such a situation, making any headway requires the interpreter to assume that many of her beliefs agree with those of the foreign speaker. So, perhaps we need to do something similar in attributing desires and intentions?

Along these lines, Donald Davidson argued that a principle of charity should be extended to the posits about what desires or values a person has. Davidson (2004b) explains the enlargement of the scope of the principle of charity this way:

> For in the plainest cases we can do no better than to interpret a sentence that a person is selectively caused to hold true by the presence of rain as meaning that it is raining... It follows that in the plainest and simple matters good interpretation will generally put interpreter and interpreted in agreement.. . Just as in coming to the best understanding I can of your beliefs I must find you coherent and correct, so I must also match up your values with mine; not, of course, in all matters, but in enough to give point to our differences. This is not, I must stress, to pretend or assume we agree. Rather, since the objects of your beliefs and values are what cause them, the only way for me to determine what those objects are is to identify objects common to us both, and take what you are caused to think and want as basically similar to what I am caused to think and want by the same objects.

This may seem to justify some leeway for workers writing descriptions to ascribe intentions to an agent depicted in some image, even when the image is consistent with several alternative claims about the agent's intentions. Perhaps we have no other way forward. But I do not think this is so. It is far from clear that this sort of charity is appropriate when the interpretive activity is not *radical* interpretation. The principle of charity is crucial when we have yet to establish that we are even talking about the same objects as the person we're interpreting. However, the principle is no longer required if enough linguistic commonality has been established that the interpreter is in a position to know the meanings of the person's sentences (or if the interpreter were in a position to ask the person for clarification). Hence, though the

kind of charity Davidson describes may be a condition of interpreting others in some unusual contexts human interaction, that does not justify allowing it as a heuristic in the generation of training data for machine learning systems. After all, relying on such a principle of charity yields presumptuous judgments. It is not exactly the same sort of presumptuousness featured in the preceding examples, but it may be just as bad. Instead of supplementing the information available with inferences based on group membership (as with stereotypes), according to the present strategy, the auxiliary information would be drawn from the inventory of mental states of the person writing the descriptions. It is no less objectionable to simply assume a person's motivations are like one's own than to assume that the person is motivated like people to whom she bears a superficial similarity.

An alternative way of attempting to resolve uncertainty about an agent's intentions is not to assume that her intentions match the interpreter's, but rather to assume that her intentions align with those that are most prevalent in the population. Daniel Dennett (1989b), working very much in the same vein as Davidson, discusses how we attribute desires when we take the so-called *intentional stance* toward an entity:

> How do we attribute the desires (preferences, goals, interests) on whose basis we will shape the list of beliefs? We attribute the desires the system *ought to have*. That is the fundamental rule. It dictates, on a first pass, that we attribute the familiar list of highest, or most basic, desires to people: survival, absence of pain, food, comfort, procreation, entertainment. Citing any one of these desires typically terminates the "Why?" game of reason giving. One is not supposed to need an ulterior motive for desiring comfort or pleasure or the prolongation of one's existence.

If indeed there are desires or intentions that are shared by all persons, then it cannot be presumptuous to judge of a particular individual that she has these desires or intentions. But notice that there is a difference between ascribing a standing desire to a person and judging that the satisfaction of that desire was the intention driving a particular action. My intention when washing the dishes is more accurately described as "getting the dishes clean" or "keeping the kitchen tidy" than in terms of any of the more basic desires Dennett mentions. So, to assume of a person that all her actions are to be interpreted as intending to satisfy these basic desires is another kind of presumptuousness.[9]

Even if an overly broad appeal to basic desires is just another form presumptuousness can take, it is worth mentioning because of the distinctive worries it raises. Some aims and values may be shared across all of humanity, but many are not. The variety among our aims is a source of richness in the human experience. To attempt to limit our interpretations of an individual's intentions to a fixed set of common goals is to underestimate the diversity of human motivations. Hence it is to view the person not as an individual with a distinctive orientation to the world, but as

[9] I do not mean to imply that Dennett himself is guilty of making this assumption.

an indistinctive node in a homogenous system.[10] Such a view, if regularly invoked, may result in the assumption of shared intentions becoming a sort of self-fulfilling prophecy, ultimately narrowing, rather than enlarging, the courses of action open to us.

Hence, it seems that neither imputing the intentions of the interpreter, nor imputing the intentions that are common, is an acceptable way to address the lack of information we have about what motivations drive the actions depicted in a photograph. The more general principle we may draw from this is that *information about one person's goals and intentions ought not be used to reach conclusions about those of another*. Again, the point here is ethical, not epistemic. It is the upshot of the preceding discussions of presumptuousness.

One possible response to this might be to argue that presumptuousness itself is not a problem. Perhaps presumptuousness is undesirable only when the intentions presumptuously ascribed appear immoral, embarrassing, or otherwise unattractive. Along these lines, suppose that, while shopping at the grocery store, I choose the expensive, environmentally friendly cleaning spray. My actual motive might be to avoid allergic reaction to a chemical in the standard variety of cleaning spray. But if people believe that my intention is to be an environmentally responsible consumer, I may not mind their inference too much. This suggests that we may not need to have image labelers withhold judgment about any and all intentions the agent may have, just the unattractive ones. The intuitive thought in the vicinity would be something like this: *It's okay to guess at persons' intentions, as long as we give the people the benefit of the doubt*. But this is not acceptable either. Although it may be a good rule of thumb for everyday social life, and although it may avoid some negative consequences of presumptuousness, it would be a totally inappropriate policy for our image recognition systems. It would bias the training data set in a way that would reduce its accuracy. After all, people often do have unattractive motives. To train the system as though this were not true would be to introduce systematic error into the system. We would be avoiding the moral problems at the expense of adding new epistemic problems.[11]

14.6 Going Beyond Still Photographic Data to Ascertain Intentions Responsibly

We have seen that attributing intentions on the basis of some kind of interpretive charity—whether the interpreter ascribes to the person the interpreter's own intentions, intentions that are common in the population, or intentions that paint the person in a positive light—is unacceptable. Supposing we still wish to develop

[10] Cf. Blum (2004).

[11] And, of course, a further worry about this strategy concerns the thorny issue about how we might go about categorizing intentions as attractive or unattractive in the first place.

systems capable of making intelligent inferences about an person's intentions, we need additional sources of data.

So, let's consider what additional data would allow responsible judgments about intentions. One limitation of the Flickr 8k data set has nothing to do with any restrictions on the descriptions the image labelers were allowed to provide. Rather the limitation is due to how the images in Flickr 8k were acquired. The researchers note that images were "manually selected to depict a variety of scenes and situations" (Hodosh et al. 2013). In effect, this means that in very few cases is any person depicted in more than one image. This fact, combined with the inherent limitations of still images, entails that the data set contains almost no diachronic information. That means that even the evidence of an agent's overt behavior is severely limited. In contrast, with several successive, timestamped photos, or with a few seconds of video, instead of just a single still image, we may have a representation of behavior sufficiently rich to ascertain more—at least something beyond the bare minimum—about an agent's intentions in acting. For instance, regarding a woman kicking a soccer ball, we might be able to say whether she is passing, shooting, or just clearing it. Unfortunately, with just a single still image we have information that is consistent with too many different possible intentions on the part of the person pictured.

Consider Fig. 14.3, which is another image from Flickr 8k.

Here are the descriptions that the Turkers wrote to describe it:

1. A man dressed for cold weather plays with a stick with his black and brown dog.
2. A man in a brown vest and glasses plays with a brown dog.
3. A man in orange pants and brown vest is playing tug-of-war with a dog.

Fig. 14.3 Another image from Flickr 8k. Note the ambiguity of the aims of the man holding the stick

4. A man tries to take a stick away from a brown dog.
5. A man tugging on a stick that a little dog has in his mouth.

All of these descriptions—except, perhaps, for (5)—display some degree of presumptuousness. Also, it is interesting to observe at least some apparent disagreement among them. While (1), (2), and (3) suggest that the man's intention is to play with the dog, (4) suggests that the man's aim is simply to get the stick. But, most importantly for present purposes, note that it would not take much additional data about this scene to make it pretty obvious which of these somewhat divergent interpretations is most correct. A few seconds of video of the scene, or a series of several photographs taken over the course of a few seconds, would likely be enough. Or, if we had a record of the man expressing a desire to play with his dog, or, alternatively, a record of him saying he intended to train his unruly canine, this might be even more helpful. This points the way to a positive recommendation, though perhaps an obvious one: *Attribution of intentions to a person, in a way that is informative, accurate, and not presumptuous, requires several data points about that particular person. Likely, the more (and the more diverse), the better.*

The task of generating training data sets that are informative, accurate, and that encode genuinely responsible judgments about persons' actions, may require using not just annotated visual information about the persons, but also data of other sorts, such as the persons' histories of verbal communication. Of course, drawing on richer data sets requires more sophisticated machine learning systems.[12] And, even with additional data about an individual, the data available may still be compatible with several different hypotheses about the person's intentions. Continuing to add more and more data about the person is the only non-presumptuous path to narrowing the set of interpretive hypotheses about a person's intentions down to just one.[13] Thus, there does, after all, appear to be a route forward that avoids presumptuousness, but it is a formidable one.

14.7 Modular Responsibility Reconsidered?

I have been considering what it would take to produce machine learning systems capable of issuing responsible judgments about the intentions with which a person acted. The approach I have considered is what I described at the outset as *ingrained*

[12] Such work is already underway. See, e.g., Park et al. (unpublished ms).

[13] Along these lines, Dennett argues, "the class of indistinguishably satisfactory models of the formal system embodied in [the] internal states [of an entity toward which we might take the intentional stance] gets smaller and smaller as we add such complexities [such as a wider range of behaviors]; the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated" (1989b). Notoriously, according to both Quine and Davidson, some indeterminacy may be ineliminable. However, along with Dennett, I doubt that any remaining indeterminacy poses any practical or ethical problems in the context of machine learning systems. For discussion of indeterminacy and its (in)significance, see Davidson (1984b).

responsibility. The thought was that we could create systems that issued only responsible judgments, by ensuring that the data on which these systems were trained included only responsible judgments. But, as we've seen, this approach will be difficult and so may require postponing benefits otherwise soon achievable.

Also at the outset I mentioned a *modular* approach as an alternative to ingrained responsibility. The idea would be to accept training data that embodies the problems, i.e., presumptuousness, that I have been discussing. And then the task would be to add an extra stage of processing that would prevent the irresponsibility from being propagated into applications. But note that an effective module for these purposes would not be just a simple filter. An algorithm that could accurately classify as presumptuous or non-presumptuous descriptions of actions may itself require machine learning. If that is so, then it seems better to avoid presumptuousness from the beginning, in the training data that might originally introduce it. In other words, it seems better to opt for ingrained responsibility.

A final worry about the modular approach is that, in practice, it may be tempting (for convenience or other reasons) to omit the extra stage of processing. The "responsibility module" might simply be left out by a developer who didn't consider it important enough to bother with. But then irresponsible judgments would make their way into computer systems we use, and we would likely never know.[14] For this reason also I hold out hope for a tractable approach to ingrained responsibility.

In light of this discussion of machine learning systems for image recognition, I venture that there is a more general—though perhaps unsurprising—lesson to be learned here: We ought to include moral criteria among the requirements for our machine learning systems and the data on which we train them, even though doing so poses distinctive and difficult challenges.

Acknowledgments I am grateful to Andréa Atkins and to attendees of the IACAP 2016 for discussion of these issues. A preliminary exposition of some of the ideas and arguments presented in this chapter appeared in a short essay posted on the website of the Loyola Center for Digital Ethics and Policy (http://www.digitalethics.org/).

References

Beeghly, Erin. 2015. What is a stereotype? What is stereotyping? *Hypatia* 30 (4): 675–691.
Blum, Lawrence. 2004. Stereotypes and stereotyping: A moral analysis. *Philosophical Papers* 33 (3): 251–289.
Davidson, Donald. 1984a. *Inquiries into truth and interpretation*. Oxford: Clarendon Press.
———. 1984b. *Belief and the basis of meaning*. Reprinted in Davidson (1984a): 141–154.
———. 2004a. *Problems of rationality*. Oxford: Clarendon Press.
———. 2004b. *Expressing evaluations*. Reprinted in Davidson (2004a): 19–37.
Dennett, Daniel. 1989a. *The intentional stance*. Cambridge, MA: MIT Press.

[14]This is a specific version of the type of problem James Moor (1985) has famously called "invisibility."

———. 1989b. *True believers*. Reprinted in Dennet (1989a): 13–35.

Fei-Fei, Li, and Li-Jia Li. 2010. What, where and who? telling the story of an image by activity classification, scene recognition and object categorization. In *Computer vision*, ed. Cipolla et al., 157–171. Berlin: Springer.

Hodosh, Micah, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47: 853–899.

Karpathy, Andrej, and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:*1412.2306.

Lippmann, Walter. 1922. *Public opinion*. New York: Macmillan.

Marvit, Moshe. 2014. How crowdworkers became the ghosts in the digital machine. *The Nation*. http://www.thenation.com/article/how-crowdworkers-became-ghosts-digital-machine/. Accessed 11 Jan 2016.

Moor, James. 1985. What is computer ethics? *Metaphilosophy* 16 (4): 266–275.

Park, Eunbyung, Xufeng Han, Tamara Berg, and Alexander Berg. (unpublishedms). *Combining multiple sources of knowledge in deep CNNs for action recognition*. http://www.cs.unc.edu/~eunbyung/papers/wacv2016_combining.pdf. Accessed 11 Jan 2016.

Quine, W.V. 1960. *Word and object*. Cambridge, MA: MIT press.

Rashtchian, Cyrus, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 139–147. Association for Computational Linguistics.

Strawson, Peter. 1962. Freedom and resentment. *Proceedings of the British Academy* 48: 1–25.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:*1411.4555.

Chapter 15
Robotic Responsibility

Anna Frammartino Wilks

Abstract This paper considers the question of whether humanoid robots may legitimately be viewed as moral agents capable of participating in the moral community. I defend the view that, in a strict sense, i.e., one informed by the fundamental criteria for moral agency, they cannot, but that they may, nonetheless, be incorporated into the moral community in another way. Specifically, I contend that they can be considered to be responsible for moral action upon an expanded view of collective responsibility, which I develop in the paper.

Keywords Moral agency · Individual moral responsibility · Collective moral responsibility · Joint commitment · Robotics · Moral community · Personhood · Autonomy

15.1 Introduction

Given the development of increasingly intelligent and *seemingly* autonomous machines, and their ongoing integration in human environments, it may not be long before we find that such machines are included, in some way, in the moral community. How such inclusion may be facilitated, and whether it is even warranted, are matters of much concern. This paper considers the question of whether highly specialized robotic beings may legitimately be viewed as moral agents capable of participating in the moral community. I defend the view that, in a strict sense, i.e., one informed by the fundamental criteria for moral agency, they may not, but that they may, perhaps, be incorporated into the moral community in a qualified sense. I contend that it may be legitimate to view certain kinds of robotic beings as members of the moral community by virtue of their capacity for a specific type of moral responsibility. I call this *forward-looking, collective, moral responsibility*,

A. F. Wilks (✉)
Department of Philosophy, Acadia University, Toronto, Canada
e-mail: anna.wilks@acadiau.ca

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_15

and provide a defense of it in the ensuing arguments. My examination of this issue is rooted in Immanuel Kant's account of moral agency and moral autonomy but is developed in connection with Margaret Gilbert's ideas of *joint commitment* and *collective responsibility*.[1] I maintain that the combination of these central moral notions yields a compelling account of how certain kinds of robotic beings may participate in the moral community, regardless of the fact that they are not genuine moral agents. Such an account would be of considerable utility in a world of increasingly greater interaction between human beings and intelligent machines.[2] Expanding our conception of moral responsibility in the manner I suggest may, I think, render appropriate the sharing of the moral sphere with certain kinds of robotic beings.

15.2 Robotic Beings and Moral Community Membership

Most of us would have difficulty denying that the very sophisticated types of intelligent machines currently being constructed by robotic engineers and artificial intelligence researchers – commonly referred to as "autonomous moral beings (AMAs)" – possess many characteristics *similar to* those possessed by moral agents.[3] We still baulk, however, at the idea of considering them members of the moral community (Shulman et al. 2009). Though there may be good reasons for refraining from accepting them as *bona fide* members of the moral community, I maintain that this does not preclude our admitting them as *qualified* members. In fact, even if they could *become* full-fledged members at some future point, I don't think this membership could happen all at once. The most rational approach is to proportion the degree of membership in the moral community to the degree of the machine's manifestation of the characteristics of genuine moral agents. Admitting intelligent machines into the domain of moral beings is something which, if it happens at all, will likely happen in stages – as happens in evolutionary processes. Hoffman (2012, 2) intimates that we have already embarked on this road, noting that "embodied robotic cognition research could transcend simple robotic systems, navigation, and dynamics, and be applied to autonomous interactive robots that act in meshed joint activities with humans." As research in artificial intelligence advances, and as increasingly sophisticated robotic beings are produced, we may

[1] Gilbert herself does not address the issue of whether this kind of responsibility may be attributed to robotic beings.

[2] Clark (2001, 31) stresses the advances being made in artificial intelligence given the new methodology adopted by many researchers in the field. This methodology focuses on constructing robotic beings that interact in the same environments as human beings.

[3] David Chalmers (2010), like Ray Kurzweil (2005) and many others, envisions a future state – the singularity – in which supremely intelligent machines *surpass* human intelligence and exercise control over the universe.

find it legitimate to incorporate these robots in the community of moral beings, incrementally. I outline, in this paper, some of the key hurdles along that path.

A crucial premise on which my general argument rests is that it may be legitimate to view a being as a member of the moral community in some senses but not others. Specifically, it may be appropriate to view the same being as having the privilege of moral rights, but not the burden of moral responsibility, and vice versa. The distinction between *moral responsibility* and *moral rights* is crucial for my position. An appreciation of this distinction is indispensable for determining the sense in which it may be possible to view certain kinds of robotic beings as capable of participating in the moral community in a meaningful way, without illegitimately granting them an unwarranted status in that community. To illustrate, consider the attitude conventionally adopted towards non-human animals, children and cognitively impaired persons. We tend to consider children as members of the moral community in the sense that they have moral rights, for example, the right to life, and the right to a certain level of well-being. We do not, however, consider them as beings capable of moral responsibility – at least not until they become able to engage in moral deliberation. We tend to believe that this capacity is only latent in them, and thus that they ought not to be held morally responsible for their actions until this capacity is sufficiently developed. In the case of cognitively impaired persons, whom we tend to view as having similar moral rights, we respect those rights even in cases where it is clear that the capacity for moral responsibility will *never* assert itself. In fact, some also consider it appropriate to view non-human animals as members of the moral community in the sense that they have moral rights that ought to be protected, unaccompanied by the burden of moral responsibility. These beliefs impose on us the obligation to protect these beings from abuse. Nonetheless, their possession of these moral rights does not entail that they are morally responsible for their behavior. Thus, for example, we do not punish non-human animals for killing other animals for sustenance.[4]

In what follows, I defend the correlative view that it may be warranted to regard certain intelligent machines as members of the moral community, in the sense that it is legitimate to ascribe to them *moral responsibility*, but not *moral rights*. The sense in which we may properly ascribe moral responsibility to them, however, is a significantly restricted one. I contend that although an otherwise sophisticated humanoid robot may not be morally autonomous (in the Kantian sense), and thus may not be considered a moral agent, we may nonetheless ascribe moral responsibility to it if it is capable of entering into a *joint commitment* with the members of a *collective* engaged in moral deliberation. The legitimacy of this position becomes apparent if the notion of *joint commitment* is understood in Gilbert's terms, i.e., as *irreducible* to the commitment of the individual members of

[4]Searle (1990, 414) maintains that, in fact, it is possible that non-human animals can form collective intentions. This issue, however, is not one that I examine here.

the collective. The kind of moral responsibility this entails is not individual moral responsibility but rather *collective moral responsibility.* I address this central issue in detail in the ensuing sections.

15.3 Collective Moral Responsibility and Joint Commitment

The account of *collective moral responsibility* operative in my argument is fueled by the view that it is possible for a collective *as a whole* to be responsible for *x*, without the members of that collective being *individually* responsible for *x*. This situation arises when individuals engage in a joint commitment to the goals of some collective. Gilbert defines *joint commitment* as: "A commitment of two or more people. It is not a conjunction of a personal commitment of one and separate personal commitments of the others. Rather, it is the commitment of them all" (Gilbert 2013, 64).[5] Gilbert places significant emphasis on the fact that a joint commitment is not reducible to the individual commitments of each of the participating parties. Rather, it is a phenomenon that results from the very *combination* of the group members.[6] As a result, "it is always the case that the parties are jointly committed to *X* as *a body*" and thus "it is to *together to constitute, as far as possible, a single body that intends to do that thing*" (Gilbert 2013, 64). In a relation of joint commitment, the "population of persons who are jointly committed in a certain way constitute a *plural subject*" (Gilbert 2013, 63). Such a population constitutes a *collective* (Gilbert 2013, 67). Gilbert stresses that a collective is not merely an aggregate, that is, a particular group the members of which share some common features, e.g., the population of persons who like opera. Gilbert conceives of a collective as a kind of *social group*, e.g., a family, team or workgroup. I employ this same notion of a collective in this paper.

In consideration of a collective consisting of a population of persons P, to which is legitimately ascribed moral responsibility for a particular state of affairs S, some crucial questions arise:

> What is the relationship of population P's moral responsibility to that of each member of P? Does P's responsibility imply, for instance, that each individual member is personally responsible for S, or at least to some extent responsible for it? Does it imply, rather, that most individual members of P are to some extent responsible for S? Or that some are? Does it imply that no individual member is to any extent personally responsible for S? Or does it have no implications either way? (Gilbert 2013, 60)

[5]Tuomela (1984) refers to this kind of intention as "we-intention." This phrase is also used by Searle (1990).

[6]Bratman (1993, 103) employs the term "shared intention" to refer to a similar notion, though Bratman distinguishes between his notion of "shared intention" and Tuomela's "we-intention," as well as Searle's "we-intending" and "collective intending."

In short, the general question here is: "What is the relationship of the moral responsibility of a collective to that of its members?" (Gilbert 2013, 60). Gilbert provides an extensive account of the sense in which it is appropriate to speak of the notion of "collective moral responsibility," and articulates a useful *model* for tracking the logical consequences of this notion. I shall be adopting the foundational concept of this model, the idea of *joint commitment*, in my account of the moral responsibility of robotic beings. In developing this notion to accommodate robotic beings, however, I stress the importance of not overstepping the line that divides genuine moral agents and mere "moral zombies." In the section that follows, I provide some direction for drawing this crucial line, to ensure that the scope of moral agency is not inappropriately extended.

15.4 Conditions for Moral Agency

It is difficult to battle the intuitive precept that *moral agency* is a fundamental criterion for moral community membership (Farah and Heberlein 2007). On the assumption that this precept can be relied upon, it would seem that if a robotic being does not satisfy the criteria for moral agency, it ought not to be considered a member of the moral community.[7] As a guiding principle, I suggest we employ the Kantian conception of moral agency in the treatment of this problem. Kant characterizes a *moral agent* as a *rational being that possesses an autonomous will* and is thereby capable of self-legislation. Self-legislation consists in the capacity to determine one's own maxims to guide one's action, as opposed to being completely determined by the laws of nature (Kant 4: 446–447). Kant asserts: "Autonomy of the will is the property of the will by which it is a law to itself" (Kant 4: 440), i.e., to be autonomous is to be *self-legislating*. Kant stresses, however, that to be self-legislating in this way involves the capacity to constrain one's choices and behavior by the moral law of reason. Only if a being possesses such capacity is it reasonable to view that being as a genuine moral agent. A being of this sort is capable of choosing the moral law both to *constrain* and to *motivate* its behavior.[8] This autonomy is what gives rational beings intrinsic worth (*dignity*), the possession of which requires that they never be treated in a merely instrumental fashion, as a means only to some end, but rather, always as *ends in themselves* (Kant 4: 432–441). According to Kant, it is autonomy of *this* kind, therefore, that is required for moral worth and for the capacity to function as a moral agent with individual moral responsibility.

[7] Farah and Heberlein (2007) draw attention to the numerous stumbling blocks for machine ethics posed by this problem.

[8] Kant defends this view in various works on practical philosophy, especially throughout *Groundwork of the Metaphysics of Morals* (1785) and *Critique of Practical Reason* (1788). All subsequent references to these texts will appear in the body of the paper, identified by the volume and pagination of the Prussian Academy Edition and refer to the translation by Guyer and Wood (1992).

Can robots be considered beings capable of functioning as moral agents in this sense?[9] Because even the most advanced current humanoid robots are designed, engineered and operated by persons with an end or set of ends in mind – commonly referred to as a *top-down* architectural model – they lack the kind of autonomy that characterizes a self-legislating being in the Kantian sense.[10] They lack, therefore, the capacity for genuine self-legislation, and can, at most, *simulate* the kind of *behavior* that self-legislating beings manifest. Simulated autonomous behavior, however, is not an adequate condition for moral agency. Thus, it would not seem that robotic beings satisfy the Kantian conditions for moral agency.

Another important feature of Kantian moral agency is that, although the moral law *requires* us to act in accordance with it, it does not *necessitate* that we do so; it merely *constrains* us to do so. What this entails is that the moral behavior of rational beings is never guaranteed; they thus remain morally fallible, since in any given instance they may or may not choose to abide by the moral law (Kant 4: 413–414). This is an essential feature of genuine *moral deliberation*. A being who is infallible in their ability to abide by the moral law would not be a rational being of the moral kind.[11] However, when artificial intelligence researchers speak of programming humanoid robots to act in accordance with moral rules, they don't seem to mean anything more than programming them to calculate, by means of some algorithm, the course of action most consistent with those rules, and then to act in accordance with that calculation. Provided the robot is in perfect functioning order, and provided there are no external constraints on it, the robot will infallibly abide by the moral rule. It does not have the capacity *not to choose* the moral law for itself.[12] Lacking this choice, however, in what way can it be said to be self-legislating, i.e., to be truly autonomous? Even if we grant that there is *reasoning* of

[9]Shulman et al. (2009) are skeptical about the appropriateness of viewing robotic beings as satisfying the criteria of moral agents.

[10]Rodney Brooks (1991a, b) points out, however, that since the 1980s there has been a notable shift in the methodology of artificial intelligence architecture. Many researchers are moving away from the Von Neuman computational models and towards behavior-based models that aim to simulate biological systems. This radical change is having a significant effect on the resulting capacities and autonomous behavior of robotic beings. Brooks expresses skepticism, however, about the plausibility of artificial intelligence researchers actually emulating the evolutionary processes at work in biological systems.

[11]It would be something like a supernatural being with a *pure will*, e.g., God; it would not be able to choose to act in a manner that contradicted the moral law.

[12]It should be noted, however, that in the new methodology of *embodied cognition,* currently adopted by researchers in artificial intelligence, robotic beings are constructed in accordance with a bottom-up architectural model as opposed to a top-down model (Hoffman 2012). These robots appear to manifest a kind of deliberation that more closely simulates the autonomous behavior of humans. If this progress continues, and there is no reason to believe that it will not, it is conceivable that some future robotic beings might be constructed that manifest genuine autonomy, as opposed to the mere simulation of autonomy. My discussion here is limited to, and only requires, consideration of the kinds of intelligent machines that have been constructed to date, since I am merely interested in establishing the *minimal* conditions for ascribing moral responsibility to robotic beings, not the maximal conditions.

some kind going on in these robots (Clark 2001), I contend that, at most, these robots can be said to be engaging in merely *instrumental* reasoning, not *moral reasoning*. Moral reasoning involves not merely the capacity to determine *what* one needs to do to behave morally, but also the *choice* to *behave morally rather than immorally*.

Moreover, for Kantian moral agency, an action's being in *accordance* with the moral law is merely a *necessary* but not a *sufficient* condition for its being morally praiseworthy. What is also required is that the action be performed from the *motive* of the moral law. That is to say, respect for the moral law must be the only incentive for the agent's acting in accordance with this law; the agent must not have ulterior motives, such as the pursuit of some end – regardless of how much utility that end may have or how honorable it may be (Kant: 4: 403–404).[13] According to Kant, an action performed in accordance with the moral law dictated by reason is a duty. Kant stresses, however, that

> an action from duty has its moral worth *not in the purpose* to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realization of the object of the action but merely upon the *principle of volition* in accordance with which the action is done without regard for any object of the faculty of desire (Kant 4: 399–400).

It is questionable, however, whether a robotic being can entertain the idea of acting from the *motive* of the moral law as distinct from merely acting in *accordance* with the moral law – as its sole aim is to bring about certain consequences. Its interest is simply the object to be achieved by some action rather than the principle motivating that action.

For these reasons, I maintain, there is no basis for viewing current robotic beings as genuine moral agents in the Kantian sense, and therefore no basis for attributing to them any kind of *individual* moral responsibility. In fact, despite the immense progress that artificial intelligence researchers and roboticists have achieved in the production of extremely sophisticated intelligent machines, not even the most advanced humanoid robots possess the criteria for moral agency as specified by Kant (Yudkowsky 2008).[14] I endorse, therefore, the claim also defended by others, that robots can, at most, only be viewed as *moral zombies* – never as full-blown *moral agents* (Howard and Muntean 2016). I shall argue, however, that this does not necessarily preclude the possibility of attributing to certain robotic beings a kind of moral responsibility that does not pertain to *individual* moral agents, but rather, to *collectives*. This is the kind of moral responsibility that Gilbert expounds in her account of joint commitment, which I examine in the section that follows.

[13] This is the central feature that distinguishes John Stuart Mill's (1979) utilitarian approach to ethics from Kant's deontological approach.

[14] Yudkowsky (2008) emphasizes the limitations of present day robotic beings and the challenges posed by attempts to include them in our moral considerations.

15.5 Can Robotic Beings Enter into Joint Commitments?

According to the foregoing analysis of the requirements for genuine moral agency, it would seem that robotic beings, regardless of how sophisticated, could not be legitimately viewed as members of the moral community, since they are not morally autonomous in the Kantian sense. I propose, however, that robotic beings may acquire some status in the moral community in a less substantial sense. Lacking the capacity for genuine moral agency, it seems clear that robotic beings would also lack the capacity to form *individual, personal commitments* – a fundamental requirement for *individual* moral responsibility. I suggest, however, that it is not inconceivable that certain kinds of robotic beings may be capable of what Gilbert refers to as *joint commitment*.

One may question the intelligibility of viewing a being as capable of sharing in *joint* commitment if they are not capable of *individual* commitment. I argue, however, that while such a claim might challenge a purely Kantian position, it seems quite consistent with Gilbert's standpoint, since, on her view, joint commitment is *irreducible* to a collection of individual commitments (Gilbert 2013).[15] Because of its irreducibility, this species of responsibility may be viewed as a genuinely novel kind of responsibility that emerges from a certain type of collective activity. This collective activity, I contend, calls for the need to *expand* the notion of responsibility to accommodate the contribution made by the collective's members. While this need manifests itself in various domains of strictly human activity, it is particularly pronounced in the domain of human-machine interaction. My aim is to show how a consideration of the insights that inform Gilbert's account of *joint commitment* can facilitate the expansion of the concept of *moral responsibility* to establish the basis for a vital aspect of machine ethics.

A crucial distinction is relevant to my account of responsibility. This is the distinction between *backward-looking responsibility* and *forward-looking responsibility*. The former concerns the causality of and accountability for past actions and the attribution of praise or blame, punishment or reward, for such actions. The latter concerns deliberation about present or future actions; it is goal or function oriented and involves the formation of commitments and the acknowledgment of obligations. My argument is intended to support only the notion of *forward-looking responsibility.* It is this kind of responsibility that most interests policy-makers in machine ethics, many of whom strongly advocate the use of intelligent machines in moral deliberation, and the moral instruction of these machines, to render them more able to contribute to such deliberation. Assessing the nature, extent, and implications of attributing forward-looking moral responsibility to intelligent machines is, therefore, a matter that calls for attention. While it *may* be possible that some aspects of my argument also have application to the issue of backward-looking responsibility in machine ethics, I do not provide a treatment of that kind of responsibility here.

[15]Gilbert (2007) notes Searle's (1990) claim that "collective intentionality" is also irreducible in this way.

My view is that because the responsibility ascribed to a collective is not one that can be ascribed to each member of that collective as an individual moral agent, it is not necessary that each member of a morally responsible collective be a moral agent in the Kantian sense. For this reason, I maintain, even a collective may, itself, be viewed as the functional equivalent of a *moral zombie* – though its *members* (or some of them) may not be moral zombies.[16] The reason is that the collective itself is not a *genuine moral agent in the Kantian sense*, and thus could not exercise *moral autonomy*. Gilbert stresses that a collective, the members of which engage in joint commitment, is not merely an *aggregate*. I agree. Nonetheless, I think that such a collective is also not a *moral agent* as characterized by Kant. This issue aside, Gilbert thinks it appropriate to attribute moral responsibility to such a collective by virtue of its being able to deliberate and act in accordance with moral values and principles that concern moral issues. In this sense, it manifests forward-looking responsibility. To warrant the attribution of this kind of forward-looking responsibility to a collective, it is sufficient, I maintain, that the collective possess a set of core features *similar* to those of moral agents, and behavior that is at least *operationally* autonomous, i.e., capable of deliberating in a rational manner about moral issues. I contend that, by the same token, it is appropriate to extend this qualified sense of *forward-looking, collective, moral responsibility* to the right kinds of robotic beings. Although it would be illegitimate to ascribe *individual* moral responsibility to a robotic being lacking the criteria for moral agency, it would *not* be unjustified to ascribe to them the kind of moral responsibility that is ascribed to a collective, provided that the robotic being is capable of entering into a joint commitment with the other members of that collective. To do this, it must be sufficiently intelligent, and must possesses the requisite degree of operational autonomy, enabling it to contribute to rational deliberation regarding the moral concerns of a collective.[17] For a robotic being to be responsible in this way is what I refer to as *robotic responsibility.*

This position rests on the premise that it is not necessary for *all* the members of a collective to be moral agents in order to ascribe moral responsibility to that collective.[18] Analogously, it is not necessary for all the members of a collective involved in the construction of a bridge to possess all the features and capacities of the other members of the collective, in order to ascribe responsibility to that collective for the integrity of the constructed bridge. Some members of this collective might be engineers, some architects, some brick layers, some mathematicians, some city planners, etc. It is not necessary that *each* member possess the same features and capacities of each of the other members in this collective. The point is that

[16]Gilbert may not agree with me on this point.

[17]These qualifying conditions preclude the possibility of extending this kind of moral responsibility to things like tables, chairs, and pencils, since they are not sufficiently intelligent, and do not possess operational autonomy. See Shani (2013) and Ziemke (1998) for further discussion of operational autonomy.

[18]The scope of this paper does not permit me to address the question of whether *any* member of the collective must be a moral agent. This, however, is an extremely complex and crucial question.

in (at least some) collectives, each member contributes something as an individual member, though the overall responsibility is not incurred by any individual member, but only by the collective – since the overall deliberation and actions taken involve the *joint commitment* of the whole collective. There is no compelling reason why the situation should change if the collective were to include a robotic being. It is not clear why it would be necessary for the robotic being to possess the same features and capacities of the moral agents in the collective, or for the moral agents to possess the same features and capacities of the robotic being in the collective. Again, each member of the collective contributes something as a unique individual member, but the responsibility for the overall task is ascribed to the group as a collective – since the ultimate goals, deliberation, and actions taken involve the joint commitment of the *collective.*

It may be objected, however, that the position I defend differs substantially from Gilbert's, since Gilbert specifically states that a joint commitment is one that holds between "two or more people" (Gilbert 2013). Thus, it seems she thinks that only *persons* may enter into joint commitments. While this may be so, it should be noted that Gilbert explicitly indicates that she is employing the term "persons" in a non-technical sense (2013, 59). She leaves the definition of this term quite open-ended. There is room, I think, in this non-specific sense of "person" to possibly include sufficiently intelligent and operationally autonomous robots. On a Kantian sense of moral agency, robotic beings that are merely operationally autonomous certainly would be excluded from the category of beings able to engage in joint commitments. Given, however, that Gilbert does not appear to adopt a strict sense of personhood and moral agency, there is nothing barring the move of considering certain kinds of robotic beings as capable of entering into joint commitments. Moreover, considering the substantially deflated sense of "commitment" that Gilbert appears to adopt, there is also no apparent requirement here for a robust sense of "intentionality" as there is for the individual commitments of genuine moral agents.

On these grounds, I argue that moral responsibility may be ascribed to robotic beings provided they manifest the capacity to engage in *joint commitments* with other members of the moral community. It is appropriate, I maintain, to view robotic beings capable of such commitments as *collectively*, though not *individually*, morally responsible for those commitments. Only when we advance beyond the attribution of *collective moral responsibility* to the attribution of *individual moral responsibility* to these robotic beings do we err.

15.6 What Is *Moral* About Robotic Responsibility?

An objection one might pose to the account I offer of robotic responsibility is that, at the end of the day, it does not turn out to be *moral* in any significant sense. After all, if the responsibility that is attributed to the collective cannot be distributed over the individual members of the collective, then it is not clear how such a collective can be held *morally responsible* for its actions any more than a rock can be held morally

responsible for falling off a cliff and killing a passerby below. In other words, what constitutes the moral dimension of the kind of responsibility that I have sketched out in this paper? This, I think, is the most challenging objection to my view. It is not, however, an insurmountable one.

The key to appreciating the sense in which the responsibility enjoyed by a collective is a moral one is to acknowledge that the responsibility derives from the *commitments* in which the members of the collective jointly engage. The falling rock that kills the innocent passerby does not engage in any commitments either to kill the passerby or not to do so. The members of the collective examined in this paper, however, *do* engage in commitments involving the carrying out of moral action. Even though, as in the case of robotic beings, those commitments follow upon the capacity for merely operational autonomy rather than genuine self-legislation, this does not completely invalidate their commitments – provided they are viewed as contributing to the joint commitment of a collective. They are only invalidated as the products of genuine intentional acts of a self-legislating moral agent.

Moreover, I maintain that while the features Kant specifies for moral agency (the ability to legislate to oneself the moral law and the ability to choose to act both in accordance with the moral law and from the motive of the moral law) *facilitate* the capacity for joint commitment, they are not *required* by it. They are only required for genuine *individual* commitment. These Kantian criteria are only necessary conditions for individual moral agency, they are not necessary conditions for joint commitment. They merely constitute *sufficient* conditions for joint commitment. The only *necessary* condition for joint commitment is the capacity for *instrumental rationality,* not *moral rationality.* Instrumental rationality is the kind of rationality that David Hume (1888) thinks is operative in morality. This kind of rationality is also at work in Mill-inspired utilitarian accounts of morality, and many contractarian moral frameworks, as well as some recently developed functionalist accounts of moral agency (Howard and Muntean 2016). The key factor in all these approaches to moral agency is the capacity for *moral calculation*, i.e., the ability to engage in moral deliberation in algorithmic fashion. It is moral decision-making *in accordance with* moral rules and principles; it is not moral decision-making *from the motive of* the moral law, as Kant demands. I take issue, therefore, with the typical characterization of Kant's deontological approach to ethics as one that simply involves "rule-following" (Wallach and Allen 2009, 84–86), given that it involves so much more. The robotic beings capable of engaging in moral deliberation of a utilitarian kind are, therefore, not genuine moral agents, but mere moral calculators and computational machines.

The question that may be posed at this point is: What kind of status in the moral community can such robotic rule-followers have? Some might argue: none at all. The fact that these robotic beings appear to *behave* like moral agents does not automatically secure their moral status over beings that do not behave like moral agents, e.g., tables, chairs, and rocks. On what basis then can the elevated moral status of extremely sophisticated robotic machines, i.e., AMA's, be justified? I suggest that assessing the legitimacy of viewing robotic beings as members capable of entering into joint commitments of the kind described by Gilbert, and thus

of properly considering them members of the moral community, may be guided by the kinds of features that Floridi and Sanders (2004) specify. These are: the reciprocal *interactivity* between the robot and its environment, the robot's *autonomy* in the sense of the ability to initiate action in a manner that is decoupled from its environment, and *adaptability*, i.e., the robot's reflective capacity to employ its own experience to alter its operating rules (Floridi and Sanders 2004). The degree to which they possess these traits, it can be argued, may be a useful guide in determining the degree to which these robotic beings may engage in *joint commitment.* The possession of such traits does *not*, I maintain, render them moral agents in the true sense, i.e., the sense that Kantian morality demands. It should be acknowledged, however, that such traits go a long way in characterizing the moral dimension of the collective responsibility that results from their joint commitments.[19]

Given how highly qualified the sense of moral responsibility is that I attribute to certain kinds of robotic beings, one may question the point of defending it. What contribution does the acknowledgment of this kind of responsibility make to the practical issues that may possibly confront us as increasingly more sophisticated robots walk off the assembly line? What practical import does this position have? Essentially, it provides a strong basis for acknowledging the appropriate kind of status that robotic beings may conceivably occupy in the moral community. In so doing, it prevents us from extending to robotic beings greater moral status than is warranted. Specifically, if restricted to a) collectives, and b) forward-looking responsibility, then robotic beings capable of joint commitment may be invested with the collective moral responsibility of engaging in moral deliberation.[20] Such moral deliberation may pertain to the domains of the workplace, the medical profession, warfare, economics, business, science and technology, care of the elderly, politics, education, etc. In short, such moral deliberation may enter into any domain that involves the overseeing of moral issues by a collective. Given the significant contributions that sophisticated robotic beings are capable of making to such moral deliberation, and increasingly will make over time, it strikes me as quite reasonable to permit them membership in the moral community in this restricted sense. I contend, however, that this status may only be enjoyed by robotic beings as *members of a collective*. It would be illegitimate to invest robotic beings with *individual* moral responsibility, given that they lack the kind of autonomy required for genuine moral agency, and the capacity to form genuine intentions and personal commitments. Investing robotic beings with this restricted kind of moral responsibility would entail granting them appropriate recognition for their capacities, without misconstruing those capacities. This involves acknowledging their contribution to moral deliberation to be greater than that of the chairs upon which the members of the collective sit, or the pens and paper they use to takes

[19] See also Ziemke (1998) for a detailed discussion of adaptive behavior in autonomous agents.

[20] Gilbert's account of collective responsibility, however, is not restricted to forward-looking responsibility.

notes. At the same time, it also involves acknowledging their contribution to be less than that which a genuine moral agent could make. The former permits the substantive use of robotic beings in collective moral deliberation; the latter guards against investing robotic beings with tasks or ranks that enable them to take on *sole* responsibility for deliberating on matters of moral import. This would be particularly pernicious if it involved granting robotic beings elevated authority over *persons* in such deliberations.

Finally, respecting these limitations on the possible inclusion of robotic beings in the moral community guards against the anticipated call for the moral rights of robotic beings that inclusion in this community may elicit. It would be a serious mistake, and potentially costly one, if, for example, we were misled into thinking that merely operationally autonomous, intelligent machines capable of engaging in rational deliberation concerning moral issues were, for that reason, entitled to the moral rights enjoyed by genuine moral agents. The restrictions I place on my general argument are, I think, sufficiently effective in averting this illegitimate move. These restrictions allow us to attribute collective *moral responsibility* to robotic beings capable of entering into joint commitments with the members of a collective that can deliberate and act on moral issues, without the necessity of granting *moral rights* to those beings. This may strike some as unintuitive since there is a strong inclination to think of moral responsibility and moral rights as inextricably linked. It seems unreasonable to ascribe to some being one but not the other. In response to this claim, I would point out that the *moral responsibility* ascribed to robotic beings capable of entering into joint commitments and deliberating on moral issues is grounded in the robot's capacity for *operational autonomy*. The operational autonomy of extremely sophisticated robotic beings (those that meet the criteria outlined by Floridi and Saunders presented above) may, I think, justify the attribution of collective, forward-looking, moral responsibility.[21] As explained, this kind of responsibility does not involve the accountability for past action that would constitute the grounds for blameworthiness or praise, and the consequent effects – punishment or reward. It involves only the capacity for moral deliberation with respect to some present or future matter. The granting of *moral rights* to these robotic beings, however, would have to be grounded in the robot's capacity for *moral autonomy* – rooted in the Kantian conception of self-legislation. However, genuine *moral autonomy* is a feature that is only *maximally* manifested by persons.[22]

[21] I do not address here the issue of whether this kind of autonomy may also justify the attribution of collective, *backward-looking* moral responsibility, though this is an important consideration.

[22] I maintain that autonomy, in the sense of self-determination, admits of degrees, and that maximal autonomy is only manifested by persons. However, lesser degrees of autonomy may be manifested by non-persons, e.g., non-human animals, or even less complex living beings, or extraterrestrial beings (if there are any). The fundamental condition for some degree of genuine autonomy is some degree of genuine (not simulated) self-determination. Even the minimal form of self-determination exhibited by the simplest living beings, in their capacity for self-organization, suffices for some minimal, i.e., non-zero, degree of genuine autonomy. I elaborate on this account of autonomy in Wilks (2016). Although this account is not one that Kant endorses, it can, I think, be rendered

Until, and unless, robotic beings are produced that are capable of at least some degree of *genuine moral autonomy*, and not just the *simulation* of this kind of autonomy, or mere operational autonomy, it is not conceivable how moral rights may properly be attributed to them. The kind of robotic beings currently being designed are very far from being able to meet such criteria. Whether we may ever find ourselves sharing a world with robotic beings that *do* meet Kant's criteria for genuine moral autonomy, and thus genuine moral agency, is not an issue that lies within the scope of this paper. My aim has only been to specify the limitations we must respect in our interactions with robotic beings that do not enjoy the status of genuine moral agents. In short, moral responsibility and moral rights are two quite distinct features of moral agency, and there is a strong need to avoid conflating them. Failure to do so may result in our mishandling the assessment of the moral status of intelligent machines in our attempts to incorporate them into the moral community.

15.7 Conclusion

I have offered a defense of the claim that the continually advancing field of artificial intelligence compels us to consider seriously the legitimacy of viewing intelligent, operationally autonomous machines as members of the moral community. I outline the special sense of *moral community membership* that I think is applicable here and argue that it is conceivable that certain kinds of robotic beings may someday meet the requirements for such membership, although they do not satisfy the criteria for genuine moral agency. Even if none of our present-day robotic beings satisfy these requirements, it is reasonable to maintain that *if* robotic beings are ever constructed that *can* satisfy these criteria, then it is legitimate to incorporate them into the moral community. Membership in this community is entailed by the appropriateness of ascribing *collective, forward-looking, moral responsibility* to these robotic beings because of their capacity to participate in *joint commitments* as members of a collective. This does not entail, however, that such robotic beings may be considered morally responsible as *individuals*.

My efforts in this paper are also directed at drawing attention to a distinction of significant magnitude typically overlooked in discussions of machine ethics. This is the distinction between the attribution of *moral responsibility* and the attribution of *moral rights*. I have argued that, especially when dealing with robotic beings, the grounds for the appropriateness of attributing one of these moral features to them are not identical to the grounds for the appropriateness of attributing the other. No adequate approach to machine ethics, I claim, can avoid serious consideration of the distinction between these central features of moral agency.

I conclude, therefore, that to prepare ourselves for the inclusion of robotic beings into the moral community, we need to adopt an expanded notion of

---

consistent with Kant's account in at least *some* respects, since Kant himself does not characterize moral agency and obligation as unique to *human persons*, but as applicable to "all rational beings" (Kant 4:425).

moral responsibility. Specifically, we need to expand this notion to accommodate the collective moral responsibility resulting from the kind of joint commitment proposed by Gilbert and constrained by the moral framework established by Kant. By appreciating the legitimacy of this kind of moral responsibility, we would, I think, be in a better position to welcome robotic beings as participants in the moral community, without mistaking them for genuine moral agents.

References

Bratman, M. 1993. Shared intention. *Ethics* 104: 97–113.

Brooks, Rodney A. 1991a. Intelligence without representation. *Artificial Intelligence* 47: 139–159.

———. 1991b. Intelligence without reason. *Computers and Thought, IJCAI-91* 1: 569–595.

Chalmers, David J. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 7: 7–65.

Clark, Andy. 2001. Reasons, robots and the extended mind. *Mind and Language* 16: 121–145.

Farah, Martha J., and Andrea S. Heberlein. 2007. Personhood and neuroscience: Naturalizing or Nihilating? *The American Journal of Bioethics* 7: 37–48.

Floridi, Luciano, and J.W. Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14: 349–379.

Gilbert, Margaret. 2007. Searle on collective intentions. In *Intentional acts and social facts*, ed. Savas Tsohatzidis, 31–48. Dordrecht: Springer.

———. 2013. *Joint commitment: How we make the social world*. Oxford: Oxford University Press.

Guyer, Paul, and Allen W. Wood, eds. 1992. *The Cambridge edition of the works of Immanuel Kant*. Cambridge: Cambridge University Press.

Hoffman, Guy. 2012. Embodied cognition for autonomous interactive robots. *Topics in Cognitive Science*: 1–14.

Howard Don, and Ioan Muntean. 2016. *A minimalist model of the artificial autonomous moral agent (AAMA)*. SSS-16 Symposium technical reports, Association for the advancement of artificial intelligence, 217–225.

Hume, David. 1888. In *A treatise of human nature*, ed. L.A. Selby-Bigge. Oxford: Clarendon Press.

Kant, Immanuel. 1900–. *Kants gesammelte Schriften.* Royal Prussian (later German) Academy of Sciences (ed.). Berlin: De Gruyter.

Kurzweil, Ray. 2005. *The singularity is near: When humans transcend biology*. New York: Viking.

Mill, John Stuart. 1979. In *Utilitarianism*, ed. George Sher. Indianapolis: Hackett.

Searle, John R. 1990. Collective intentions and actions. In *Intentions in communication*, ed. P. Cohen, J. Morgan, and M.E. Pollack, 401–415. Cambridge, MA: MIT Press.

Shani, Itay. 2013. Setting the bar for cognitive agency: Or, how minimally autonomous can an autonomous agent be? *New Ideas in Psychology* 31: 151–165.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. 2009. Machine ethics and superintelligence. In *AP-CAP 2009: Proceedings of the fifth Asia-Pacific computing and philosophy conference,* ed. Carson Reynolds and Alvaro Cassinelli, 95–97.

Tuomela, R. 1984. *A theory of social action*. Dordrecht: Reidel.

Wallach, Wendell, and Colin Allen. 2009. *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Wilks, Anna F. 2016. Kantian foundations for a cosmocentric ethic. In *The ethics of space exploration*, ed. James S.J. Schwartz and Tony Milligan, 181–194. Cham: Springer.

Yudkowsky, Eliezer. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks*, ed. Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

Ziemke, T. 1998. Adaptive behavior in autonomous agents. *Presence* 7: 564–587.

Chapter 16
Robots, Ethics, and Intimacy: The Need for Scientific Research

Jason Borenstein and Ronald Arkin

Abstract Intimate relationships between robots and human beings may begin to form in the near future. Market forces, customer demand, and other factors may drive the creation of various forms of robots to which humans may form strong emotional attachments. Yet prior to the technology becoming fully actualized, numerous ethical, legal, and social issues must be addressed. This could be accomplished in part by establishing a rigorous scientific research agenda in the realm of intimate robotics, the aim of which would be to explore what effects the technology may have on users and on society more generally. Our goal is not to resolve whether the development of intimate robots is ethically appropriate. Rather, we contend that if such robots are going to be designed, then an obligation emerges to prevent harm that the technology could cause.

Keywords Social robotics · Intimate relationships · Affective bonding · Embodied robots · Well-being

16.1 Introduction

Ethical concerns about robotic technology have garnered much attention, especially in the context of how it may be used for military engagements. Understandably, there is much trepidation about whether, and in which circumstances, robots should be used in war. Although perhaps not as ethically weighty as their use in the military context, an emerging and significant area of concern is how robotic technology could affect the well-being of humans during the course of their daily lives. More specifically, intimate relationships will begin to form in the near future between robots and human beings. It is the sort of development that should not be treated

J. Borenstein (✉) · R. Arkin
Center for Ethics and Technology, School of Public Policy, Ivan Allen College of Liberal Arts, Atlanta, GA, USA
e-mail: jason.borenstein@pubpolicy.gatech.edu

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_16

lightly; it certainly deserves thorough ethical scrutiny. The advent of situated, embodied, and responsive robotic technology can have a profound impact on the social fabric of communities if and when people start to truly care about and form loving attachments to robotic artifacts. What may mitigate some of the concern about intimate robotics is ensuring that the realm is informed by rigorous scientific research, which systematically examines the associated ethical, legal, and social issues. Correspondingly, this paper will focus on two key aspects of intimate robotics. First, we will seek to identify key ethical concerns associated with this technological realm. Second, we will articulate some of the main research questions that need to be addressed prior to intimate robotics becoming a reality.

16.1.1 Defining Intimacy

Intimacy can be difficult to define precisely, but it undeniably encompasses thoughts and behaviors beyond those merely involving physical sex acts. In many circumstances, it refers to interactions that do not have a sexual dimension. The implication is that if a relationship is intimate, it contains a facet of strong emotional attachment or even love. Although "love" is a notoriously difficult to define term, many authors shed light on that term may mean by categorizing its different types (e.g., Sullins 2012, 401; Sternberg 1986). For our purposes, we operate with the assumption that love, in a broad and encompassing sense of the term, is an essential component of intimate relationships.

Technological artifacts can play a key role in the formation of intimate relationships between human beings. Within that context, intimacy can encompass (Bell et al. 2003):

- Cognitive and emotional closeness with technology, where the technology may be aware of and responsive to our intentions, actions, and feelings.
- Physical closeness with technology, either on or within the body.
- The use of technology to express our intentions, emotions, and feelings towards others.

In some ways, it is natural for human to form affective bonds with animate and inanimate non-human entities. Young children, for example, seem rather predisposed to form strong attachments to items such as blankets and toys. And certainly both children and adults form emotional bonds with their pets (Levy 2007, 46–63). More specifically related to technological devices, Reeves and Nass state (1996):

> Equating mediated and real life is neither rare nor unreasonable. It is very common, it is easy to foster, it does not depend on fancy media equipment, and thinking will not make it go away. . . . Media equal[s] real life applies to everyone, it applies often, and it is highly consequential. And this is surprising.

Commonly expressed predictions about the near future suggest that the design of robots will progress in such a way that the technology will effectively be able to establish intimate relationships with a broad range of human beings. Of course, users have already bonded with robots to some degree (for example, children developing feelings of affection for *Paro* or *Keepon*). But the intensity and scale of these attachments are expected to change dramatically with the increasing sophistication of robots.

16.1.2 Intimate Robotics

Many types of robots are in the process of being developed but the discussion here will largely focus on "human-like" robots designed to serve as companions for people. At least some of these robots may eventually have an intimate relationship with a human being, which could include a sexual component. Fictional robots that display sexual features go back at least to the 1927 movie *Metropolis* (Perkowitz 2004, 27–29). In 2007, Levy examined the technological state of the art in sexual robotics. He and other scholars describe scenarios where humans might seek out a robot in order to satisfy their physical desires. However, intimate robotics is fundamentally different as it does not simply include physical sex, which up until now has been the realm of science fiction (e.g., Blade Runner, Cherry 2000, AI, Asimov's Robots of Dawn, Battlestar Galactica, Data from Star Trek: The Next Generation, etc.). Robots could be designed in ways that move beyond being just involved in sex acts and yet are still considered intimate in a broad sense of the term.

Sullins (2012, 298) is correct that the robotics community is not yet near creating an android that is indistinguishable from a human; yet that is not a strict requirement for a human being to have intimate feelings for a robot. What makes robots unique in terms of the types of intimate relationships they may be able to form with a human is the sophistication of the traits they can possess. Physical robots can display behaviors capable of inducing feelings of attachment from human users (Bowlby 1979). This can be accomplished through a variety of methods, including affective modeling (Moshkina et al. 2011), behavior generation (Arkin et al. 2003), kinesics, haptics, and proxemics (Brooks and Arkin 2007), which may yield significant unidirectional affective bond formation between a human and the robot. As opposed to the visual and auditory experiences that computing devices can already provide, a physical robot could in addition hold someone's hand or pat one's shoulder. Or it could provide emotional support through a hug or verbal discourse. Furthermore, some roboticists (e.g., Samani et al. 2011) are specifically examining the design features of a robot that may lead to the formation of mutual love between it and a human.

The embodiment of robots raises the stakes with respect to love and intimacy as opposed to merely sexual objects. This is a key part of the reason why ethical and social issues related to robots are more complex, and perhaps more troubling, than

those associated with sex machines or toys. The feeling of intimacy and bonding with a robot as a result of persistence and embodiment, and not just physical sex, can become a likely reality. Human users could certainly believe as though they are in an intimate relationship with a robot without the robot genuinely reciprocating the feelings directed towards it.

16.1.3 Effects on the User

While the justification for the development of innovative technology, including robotics, is often couched in the language of liberty maximization, scholars have warned about the potential deleterious effects of the bonding that can form as a result of human-robot interaction (HRI) (e.g., Scheutz 2012); the ethical ramifications of these bonds have been explored in some depth (e.g., Sparrow 2002). Similar issues can also emerge within the context of elder and child care if a relationship of trust is established (Sharkey and Sharkey 2010; Borenstein and Pearson 2010).

As alluded to above, intimate robotics is fundamentally different from sex toys and devices which have been with human beings for millennia. The risks that a robot may uniquely pose are related to its embodiment (as different from pornographic videos or games), situatedness (being collocated with the human partner contextually), affective attachment, and responsiveness (as different from sexual paraphernalia). User expectations for a robot will be molded by the technology's similarities in appearance and behavior to a human. The bonding between a robot and a human partner may lead to unprecedented changes in society that are difficult to foresee although some scholars have sought to articulate the associated risks (Sparrow 2002).

Among the key ethical issues that warrant examination from the perspective of how the user might be affected include whether, and in which ways, intimate robotics may uphold or erode autonomy. The case could arguably be made that the technology supports autonomous decision making by allowing the user to select from a range of relationships options. Those who have difficulty forming social bonds, perhaps in some cases due to bad experiences, shyness, or a disability, might prefer to interact with technology. The existence of a companion robot could be viewed as offering the user as an alternative to what may be perceived as difficult or emotionally taxing situations.

However, a common concern, often discussed in the context of healthcare, is whether introducing robots into a user's life might constitute a form of deception (Sparrow and Sparrow 2006); the user may project traits or characteristics onto the technology that it does not possess (e.g., the robot "cares" about me or is "happy" to see me) (Borenstein and Pearson 2013, 184–186). There are even reports of U.S. military personnel forming attachments with bomb disposal robots (Michel 2013). Humans certainly have a psychological predilection for anthropomorphizing pets and other entities, which can lead to the formation of a powerful emotional bond (Levy 2007). At times, generating this type of user response is what a roboticist

deliberately intended; it may have resulted from a series of calculated design decisions (Arkin et al. 2012). As Sullins notes, given the roboticist's ability to design technology that elicits strong emotional responses from a user, it may follow that human-robot relationships can be established which are "as real and moving as those we have with our beloved pets and insincere lovers" (2012, 399). While we do not need to be committed to a point of view on whether a robot will be able to genuinely love a human being, a user could plausibly "fall in love" with a robot. Some users seem to already have feelings of love, at least in some sense of the term, for technological artifacts (Levy 2007).

The user may experience a range of psychological effects while interacting with an intimate robot, some of which may be rather difficult to predict. For example, how might a "risk-free" relationship with a robot affect the mental and social development of a user? Presumably, a robot would not be programmed to break up with a human companion; and thus, theoretically, this would result in the removal of the emotional highs and lows from the relationship. A similar concern has been articulated in the context of the formation of connections online where some individuals may call each other "friends" but have never had met one another (Dreyfus 2004, 77–78). For example, they may have temporary interactions, such as playing online games together, but not necessarily have to navigate through the full range of challenges that can be associated with friendship. Yet what has been learned from empirical studies of online friendships will not necessarily map directly onto what may occur in the context of HRI. The lack of an ability for the robot to rebuff the human user may, for example, lead to a behavioral deviation from the human norm that may push the user into the uncanny valley (Mori 2012) where the artifact becomes substantially less satisfying and realistic; this could at a minimum disrupt the illusion of willing participation.

16.1.4 Altering Human–Human Interaction

As Turkle states, "A relationship with a computer can influence people's conception of themselves, their jobs, their relationships with other people, and with their ways of thinking about social processes" (1984, 168). Intimate human-robot partnerships may have a similar impact on human-human relationships and on society more broadly. Associated concerns include the effects intimate robotics may have with respect to the stability of marital, pre-marital, and courtship relationships. For instance, feelings of jealousy may emerge from the amount of time that a significant other spends with a robot. On the other hand, the technology could be used to enhance the intimacy that couples experience with one another. The loss of contact with fellow humans and perhaps the withdrawal from normal everyday relationships is also a possibility. For example, a user who has a companion robot may be reluctant to go to events (e.g., a wedding) where the typical social convention is to attend as a couple.

Moreover, intense stigmatization may occur in response to intimate human-robot relationships; it is not outlandish to predict that humans in these relationships might fear for their safety given how human society often persecutes those who are perceived to be "abnormal"; at least some religious perspectives are likely to consider a robot-human relationship to be "sinful" and perhaps something that warrants punishment.

Even changes in the workforce in terms of new job creation and the effect on human performance at their existing jobs would not be unexpected should some form of addiction to these artifacts manifest itself. Given the potential for companion robots to alter the nature of human-human relationships and even the definition of love, we suggest that this realm warrants more extensive research and, to echo Whitby's sentiment (2012, 243), greater public scrutiny.

16.1.5 The Status of Intimate Robotics as a Research Field

Social and companion robotics is currently a highly active research field, with numerous conferences on the subject. Rarely, however, is the subject of intimate HRI broached in serious scholarly venues; this is largely because the realm is still considered taboo. There are some attempts to rectify this, notably the series of conferences on Love and Sex with Robots.[1] Nonetheless, robotic artifacts that are more or less sexual devices are being developed and marketed in a technically unsophisticated manner. Sex and pornography played a large role in the development of video recording devices and the Internet, and robotics is probably not immune from that type of influence.[2] Moreover, the realm of intimate robotics, as mentioned previously, is not just about sexual devices; it is about a broad category of technology with which human users might form strong emotional attachments.

Drawing from Sternberg's Triangular Theory of Love (1986), it is worth investigating whether the three key components of love (intimacy, passion, and commitment) will form between humans and robots. The empirical findings from psychologists and others about the courtship behaviors that facilitate bonding (e.g., Renninger et al. 2004; Grammer 1989) will likely influence roboticists as they design intimate robots. This type of strategic effort could significantly affect, and potentially harm, users in numerous ways.

[1] Refer to http://loveandsexwithrobots.org/. Accessed 14 Sept 2015.

[2] Don Norman, personal communication with one of the co-authors.

16.1.6 Establishing a Research Agenda

An argument in the engineering world that often emerges with regard to the development of an ethically contentious technology (including weapon systems) is that if the technology is not created by "us", then someone else will inevitably do so. While it may be ethically dubious to create atomic weapons, for example, the refusal to pursue their development may put a nation at a serious disadvantage. This is hardly to say that the point of view is necessarily correct; there are many critics of this type of argument in part because it may be used as a strategy to try to morally insulate designers from blame or accountability for the technology that they create. A similar type of argument, and resulting counterarguments, can be voiced with regard to intimate robots. We will not seek to resolve the dispute here about whether creating the technology is ethically acceptable. However, consumer demand and market forces will likely drive the development of intimate robots, and if this is the case and the effects of the technology are not rigorously studied, there will be much potential for users and others to experience profound harm.

If we operate with the assumption that intimate robot-human relationships are going to become reality, then an ethical imperative to develop a comprehensive scientific research agenda emerges, which seeks answers to questions that could prevent harm to users and others. Of course, there is the overarching issue of whether certain types of research questions in this realm are unethical or otherwise inappropriate to explore but that is not something we seek to resolve here. Rather, our purpose is to identify key research questions that should be addressed prior to allowing intimate robots to become pervasive.

Among the research questions that warrant exploration include:

- Which kinds of beliefs and attitudes about intimate robots are likely to emerge from users who interact with the technology?

Users may view these robots as just another form of technological artifact (like how a computer is standardly perceived) or alternatively more meaningful emotional attachments might form. Turkle notes that children tend to think toys that move are "alive" (2006, 8); furthermore, children seem to grieve when electronic devices like the Tamagotchi "die" (2011, 33–34 & 42–44). Is this psychological reaction likely to carry over to adults if and when robots appear to behave in more sophisticated ways? Based on the interactions humans have with non-human entities, Levy (2007) makes a compelling case that people will likely form strong emotional attachments to robots and even fall in love with them. Users will likely draw on past experience with other humans as a reference point for forming expectations about how a robot might behave (Feil-Seifer and Mataric 2011, 27) and perhaps for how the robot "feels" about them.

- How might intimate robots contribute to, or fail to contribute to, the well-being of users?

If users sincerely believe that they are in a loving relationship with a robot, will they experience benefits that are similar to being part of a human couple? Humans have various emotional and other needs that drive them to seek out companionship; these needs include self-esteem, having a sense of affiliation, and self-actualization (Sternberg 1986, 122). Humans already trust, and arguably overtrust (Carr 2014), many electronic devices including computers, GPS, and smart phones. However, as compared to other devices, a user's identity and well-being may be more integrally tied an intimate robot; the technology can pose rather unique risks to a user if emotional attachment and feelings of love emerge.

- Would the use of technology change beliefs, attitudes, and/or values related to human-human relationships and if so, how?

As mentioned previously, concern persists about whether the technology might disrupt human relationships such as marriages (e.g., a scenario displayed in the fictional TV show *Humans*). It is an open question whether users may become less tolerant of human idiosyncrasies and failings; perhaps some will become impatient and become unwilling to put the effort into working on human-human relationships. Moreover, some humans seek out prostitutes or other non-traditional arrangements due to the "lack of complications" (Levy 2007, 210); presumably, this could carry over to individuals who prefer a relationship with a robot and avoid the challenges associated with intimacy formation between human beings.

Another facet of the topic is if one's significant other has an intimate interaction with a robot, does this constitute a form of cheating? For example, some couples may live in different cities from one another and perhaps will desire the companionship of a robot while the other person is away. Given that there are different perceptions on whether "cyberromance" should be characterized as being unfaithful (Levy 2007, 45), it is safe to assume that couples will disagree on this matter.

- Are there different cultural or religious perceptions of what is appropriate practice in this realm, and if so, what might that mean in terms of societal acceptance or rejection of the technology?

Generalizations are widespread about how different cultures perceive robotic technology (e.g., that Japanese people are technophiles and that American society has deep-seated fears about robots (Kaplan 2004)). Indeed, robots have already served as officiants in Japanese wedding ceremonies (I-Fairy Robot 2010) and have even been married to each other (McCormack 2015). Perkowitz notes that Japanese religions are often seen by scholars as not drawing a sharp distinction between animate and non-animate beings, while "Western religion is hostile to artificial beings, the creation of which is seen as impious or worse" (2004, 215–216). In fact, some efforts are forming to ban the practice of robot prostitution (Brown 2015). Investigating the alleged differences in cultural and religious perceptions of intimate robots could be an important facet of a larger research endeavor.

- Is there a (causal) connection between interacting with an intimate robot and an increase or decrease in violent behavior by the user of the technology?

For many reasons, this would be a difficult research question to resolve but it is crucial to unearth any relevant data. With regard to debates about the morality of pornography, there are views all along the spectrum on the effects it may have on viewers. On one end of the spectrum, some argue that it may be an outlet for sexual desire and thus reduces the likelihood of violence; on the other end of the spectrum, one might contend that it distorts perceptions about the value of a human being and may intensify the desire to harm others. Similarly, the causative and/or correlative connections between interacting with an intimate robot and the effect on user behavior would need to be investigated. For example, if a user repeatedly kicks a companion robot, one could ask whether there is anything unethical about such acts. Kate Darling at MIT has investigated this to some degree but in the context of "torturing" a robot (Lalji 2015; Daily Mail Reporter 2013). Yet what is likely to be of greater concern, however, is whether this would establish a pattern of behavior that may eventually affect other humans. If a robot's sensors are not advanced enough to distinguish between a tap and a malicious kick, it could reinforce bad behavior by the user. In short, would the normalization of consequence-free violence in the user's personal life eventually affect other people?

- Could an intimate robot serve a therapeutic purpose for certain kinds of medical or sociopathic conditions?

One could imagine that a person who was a victim of a traumatic event (such as a physical assault) would naturally have difficulty trusting other people. Arguably, a robot who befriends a traumatized person might be viewed as a soothing intervention. On the other hand, some might interpret the strategy as being highly insulting and insensitive. On a different note, Levy suggests that robots may be useful for those "who suffer from psychosexual hang-ups" (2007, 308); a fictional example similar to this is displayed in the 2007 movie *Lars and the Real Girl* where the main character has a romantic relationship with a doll.[3] Another possible goal for researchers is to identify interventions involving robotic technology that could reduce rates of recidivism among those who commit sex crimes.

16.1.7 Conclusion

A vast array of intimate robots is seemingly on the horizon. If roboticists intend to continue pursuing this design pathway, the technology that they build could significantly impact the well-being of users and the stability of human-human relationships. Given this state of affairs, it entails an ethical obligation to systematically investigate the likely effects that the technology may have on society. Although

[3]Refer to http://www.imdb.com/title/tt0805564/. Accessed 21 Sept 2015.

many may consider this realm of inquiry taboo, the overarching aim of preventing harm to users and their communities is one worthy of pursuit and actually may invoke a moral imperative to do so.

References

Arkin, R., M. Fujita, T. Takagi, and R. Hasegawa. 2003. An ethological and emotional basis for human-robot interaction. *Robotics and Autonomous Systems* 42 (3–4): 191–201.

Arkin, R.C., P. Ulam, and A.R. Wagner. 2012. Moral decision-making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE* 100 (3): 571–589.

Bell, G., T. Brooke, E. Churchhill, and E. Paulos. 2003. *Intimate ubiquitous computing*. Proceedings UbiComp.

Borenstein, J., and Y. Pearson. 2010. Robot caregivers: Harbingers of expanded freedom for all? *Ethics and Information Technology* 12 (3): 277–288.

———. 2013. Companion robots and the emotional development of children. *Law, Innovation and Technology* 5 (2): 172–189.

Bowlby, J. 1979. *The making and breaking of affectional bonds*. London: Tavistock Publications.

Brooks, A., and R.C. Arkin. 2007. Behavioral overlays for non-verbal communication expression on a humanoid robot. *Autonomous Robots* 22 (1): 55–75.

Brown, E.N. 2015. Campaign against sex robots launches, because some people will panic about anything. Reason.com, Sept 15, https://reason.com/blog/2015/09/15/campaign-against-sex-robots-launches. Accessed 11 Aug 2016.

Carr, N. 2014. *The glass cage: Automation and us*. New York: W.W. Norton & Company.

Daily Mail Reporter. 2013. The rise of the machines: It is okay to torture a robot? (sic). *The Daily Mail*, Nov 29, http://www.dailymail.co.uk/news/article-2515400/The-rise-machines-It-okay-torture-robot.html. Accessed 11 Aug 2016.

Dreyfus, H. 2004. Nihilism on the information highway: Anonymity versus commitment in the present age. In *Community in the digital age: Philosophy and practice*, ed. A. Feenberg and D. Barney, 69–81. Lanham: Rowman & Littlefield.

Feil-Seifer, D., and M.J. Mataric. 2011. Socially assistive robotics: Ethical issues related to technology. *IEEE: Robotics and Automation* 18 (1): 24–31.

Grammer, K. 1989. Human courtship behaviour: Biological basis and cognitive processing. In *The sociobiology of sexual and reproductive strategies*, ed. A.E. Rasa, C. Vogel, and E. Voland. New York: Chapmann and Hall.

I-Fairy robot weds Japanese couple. 2010. *The Guardian,* May 16, http://www.theguardian.com/technology/2010/may/16/ifairy-robot-wedding-japan. Accessed 11 Aug 2016.

Kaplan, F. 2004. Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics* 1 (3): 1–16.

Lalji, N. 2015. Can we learn about empathy from torturing robots? This MIT researcher is giving it a try. *YES! Magazine*, July 14, http://www.yesmagazine.org/happiness/should-we-be-kind-to-robots-kate-darling. Accessed 11 Aug 2016.

Levy, D. 2007. *Love and sex with robots*. London: Harper Perennial.

McCormack, S. 2015. Robots are getting married now. *The Huffington Post*, June 30, http://www.huffingtonpost.com/2015/06/30/robot-wedding_n_7696666.html. Accessed 11 Aug 2016.

Michel, A.H. 2013. *Interview: The professor of robot love. Center for the Study of the Drone*. Oct 25, http://dronecenter.bard.edu/interview-professor-robot-love/. Accessed 11 Aug 2016.

Mori, M. 2012. The uncanny valley. Translated by K.F. MacDorman and N. Kageki. *IEEE Spectrum*, June 12, http://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley. Accessed 11 Aug 2016.

Moshkina, L., S. Park, R.C. Arkin, J.K. Lee, and H. Jung. 2011. TAME: Time-Varying affective response for humanoid robots. *International Journal of Social Robotics* 3 (3): 207–221.

Perkowitz, S. 2004. *Digital people: From bionic humans to androids*. Washington, DC: Joseph Henry Press.

Reeves, B., and C. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Stanford: The Center for the Study of Language and Information Publications.

Renninger, L.A., T.J. Wade, and K. Grammer. 2004. Getting that female glance patterns and consequences of male nonverbal behavior in courtship contexts. *Evolution and Human Behavior* 25: 416–431.

Samani, H.A., A.D. Cheok, M.J. Tharakan, J. Koh, and N. Fernando. 2011. *A design process for lovotics*. Proceedings of the International Conference on Human-Robot Personal Relationships, 118–125.

Scheutz, M. 2012. The inherent dangers of unidirectional emotional bonds between humans and social robots. In *Robot ethics: The ethical and social implications of robotics*, ed. P. Lin, K. Abney, and G. Bekey, 205–221. Cambridge, MA: MIT Press.

Sharkey, N., and A. Sharkey. 2010. The crying shame of robot nannies: An ethical appraisal. *Interaction Studies* 11 (2): 161–190.

Sparrow, R. 2002. The march of the robot dogs. *Ethics and Information Technology* 4 (4): 305–318.

Sparrow, R., and L. Sparrow. 2006. In the hands of machines? The future of aged care. *Minds and Machines* 16: 141–161.

Sternberg, R.J. 1986. A triangular theory of love. *Psychological Review* 93: 119–135.

Sullins, J. 2012. Robots, love, and sex: The ethics of building a love machine. *IEEE Transactions on Affective Computing* 3 (4): 398–409.

Turkle, S. 1984. *The second self: Computers and the human spirit*. New York: Simon and Schuster.

———. 2006. *A nascent robotics culture: New complicities for companionship*, AAAI technical report series. Palo Alto: Routledge.

———. 2011. *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.

Whitby, B. 2012. Do you want a robot lover? The ethics of caring technologies. In *Robot ethics: The ethical and social implications of robotics*, ed. P. Lin, K. Abney, and G. Bekey, 233–248. Cambridge, MA: MIT Press.

Chapter 17
Applying a Social-Relational Model to Explore the Curious Case of hitchBOT

Frances Grodzinsky, Marty J. Wolf, and Keith Miller

Abstract This paper applies social-relational models of moral standing of robots to cases where the encounters between the robot and humans are relatively brief. Our analysis spans the spectrum of non-social robots to fully-social robots. We consider cases where the encounters are between a stranger and the robot and do not include its owner or operator. We conclude that the developers of robots that might be encountered by other people when the owner is not present cannot wash their hands of responsibility. They must take care with how they develop the robot's interface with people and take into account how that interface influences the social relationship between it and people, and, thus, the moral standing of the robot with each person it encounters. Furthermore, we claim that developers have responsibility for the impact social robots have on the quality of human social relationships.

Keywords HitchBOT · Robot-human interaction · Robotic interfaces · Social robotics · Social-relational model · Anthropomorphic framing · Robotic design

17.1 Introduction

HitchBOT, a robot under the direction of David Smith and Frauke Zeller was designed as a "free-spirited robot who[1] wants to explore the world and meet new friends along the way" (hitchbot 2015). It began its exploration in 2014, hitchhiking

[1]The hitchBOT handlers refer to their robot using "who," a pronoun typically reserved for humans. Further, ascribing the act of developing friendships is typically not ascribed to robots. These issues are taken up later in the paper.

F. Grodzinsky (✉)
Hersher Institute, Sacred Heart University, Fairfield, CT, USA
e-mail: grodzinskyf@sacredheart.edu

M. J. Wolf
Bemidji State University, Bemidji, MN, USA

K. Miller
University of Missouri, St. Louis, MO, USA

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_17

across Canada, Germany, and the Netherlands. In July 2015, hitchBOT began an American adventure, intending to hitchhike from Massachusetts to San Francisco. HitchBOT had a bucket list, created with the help of "friends" on its website, and could be followed in Twitter.

On August 1, 2015, while waiting for its next ride in Philadelphia's Old City, hitchBOT was destroyed. According to Smith, "As researchers, we wanted to know, 'can robots trust humans?' and knew there would always be the possibility that hitchBOT would be damaged or stolen" (VanderMaas 2015). HitchBOT fans were outraged. According to the press release that followed, they offered condolences to the creators and developers, offered to search for hitchBOT's parts and send them back to Canada, and contributed money for another iteration of the robot. This did not come as a surprise to hitchBOT's creators.

The question posed by its creators – "can robots trust humans?" – assumes that robots can participate in authentic face-to-face trust relationships with humans. The homepage for the project has a "quote" from hitchBOT that suggests that the robot is capable of emotion: "My love for humans will never fade." Some may object to that question and that quote as inappropriate for a machine. Others may counter that the website was meant to be tongue in cheek; the website has a whimsical flavor, but we contend that even if a website is meant to be casual and fun, using such words and referring to such ideas matter.

The trust that hitchBOT's creators suggest is a face-to-face AA (artificial agent) → H (human) relationship as classified in Grodzinsky et al. (2011). We do not claim that an artificial agent such as hitchBOT would experience trust in a way identical to humans. However, the behavior of the robot might appear to be analogous to what an observer might expect in a human-to-human trust relationship. The social experiment of hitchBOT allowed people to pick it up and give it rides. That hitchBOT "trusted" them and demonstrated that trust by going with strangers, is (we contend) an anthropomorphization in the minds of those interacting with hitchBOT. They *related* to hitchBOT and *responded* to it when it seemed friendly and unthreatening, and when it acknowledged cues. HitchBOT *seemed to* trust the humans. This is an example of what Kate Darling calls "anthropomorphic framing," that is "introducing robotic technology through anthropomorphic terminology and narrative" (Darling 2015, 1). Yet despite terms such as "family" and "bucket list" used on the hitchBOT website, there is no indication that the developers actually tried to program into the robot the kinds of emotions and intuitions necessary for a human trust relationship; that kind of programming would be a major advance over published artificial intelligence (AI) scholarship, and we have no indications of such sophistication in hitchBOT. So we contend that any "trusting" going on "by hitchBOT" would be quite dissimilar to the kind of trust we traditionally expect from humans. A more critical view is that the developers of hitchBOT are misleading, essentially breaching the trust between themselves and people who interact with hitchBOT by giving hitchBOT superficial characteristics that deceive humans into believing that hitchBOT is capable of human-like trust (see Grodzinsky et al. 2015).

The hitchBOT design encourages people to believe that hitchBOT "feels" certain things about humans in general and about certain people more specifically, soliciting

empathy. Studies by Turkle (2011) and Darling reveal that "...compared to virtual presence, the physical presence of a robot affects unconscious human perception of the robot as a social partner, including self-reported empathy" (Darling 2015, 3). This might help explain the outcry when hitchBOT was destroyed.

Currently, most interactions with robots are job specific. Some robots are used in situations that are too dangerous for humans (manufacturing, mining, military applications), and others for mundane tasks such as cleaning our houses (Roomba). According to Darling, these robots should not be designed to mimic anthropomorphic traits. Social robots such as eldercare robots or therapeutic pets are more effective when they display human-like traits because it is easier for people to view such robots as social partners. These human-like traits could be considered benign deceptions (see Grodzinsky et al. 2015).

Should we consider hitchBOT a benign deception? It waited for rides and had limited interactions with humans. Its actions mimicked (in some form) what people expect from a human hitchhiker. It was built to become part of the sociotechnical system of road transportation. So, what differentiated the human reaction to hitchBOT as a cute, robotic hitchhiker, from an equally human reaction: something to be destroyed? Was hitchBOT's destruction malicious mischief, or was it something more profound? What does the empathy or lack of empathy toward hitchBOT say, if anything, about the robot's moral status? Motivated by the hitchBOT story, we explore Mark Coeckelbergh's social-relational model and Darling's notion of anthropomorphic framing as they apply to robots. We stretch the models to consider a range of robots based on how social they are and identify weaknesses when the social relationship is a casual encounter when the robot owner is not present. We analyze issues raised by these models not only from interactions between people and robots, but also from the point of view of developers of the interface of the robot, whose choices impact how people experience a robot. Exploring these questions can carry us forward in a more nuanced understanding of how we can, and should, relate to robots in society.

17.2 Social-Relational Models for Considering a Robot's Moral Status

Coeckelbergh proposes a social-relational model for considering a robot's moral status (2010, 2014). He is critical of standard approaches of determining whether robots deserve some sort of moral standing since those approaches demand that one "should investigate if the entity in question has the morally relevant property" (2014, 62). He identifies two epistemological problems for those approaches: (1) How can we know if an entity has that property? and (2) How do we know that a having a particular property makes an entity deserve a moral status (2014, 63)? He then argues that even if these two problems could be solved, we still cannot solve what he calls a gap between reasoning and experience: even when people know intellectually

and analytically that they are dealing with a machine, and that they cannot determine its moral standing in any formal way, in day-to-day interactions with robots, humans tend to "treat the robot as if it has human or animal properties—including moral status" (2014, 62). Their *experiences* with the robot are not limited to their analysis of the robot's formally established moral status. Coeckelbergh's observation seems to us to be potentially useful as we consider hitchBOT and how people related to this particular robot. We contend that relationships will help us explore the outpouring of support, sympathy, and outrage that came with hitchBOT's demise. Relationships may also help us think about why someone destroyed the robot.

Darling identifies social robots as social partners because their benefits "are most effectively achieved through a social relationship between the human and the robot" (2015, 1). She approaches the idea of relationships between humans and their social robots in a way that seems to us to be largely consistent with Coeckelbergh. Where she differs from Coeckelbergh is that she does not address the moral status of robots, but is concerned with how anthropomorphic framing can influence robot design and influence policy and law surrounding robots. Darling's social robots come with a narrative that exists prior to interaction, a narrative that does not change because of contacts between humans and the robot. Coeckelbergh, on the other hand, seems to suggest that constituting the moral status of a robot is a dynamic process, one that varies with human subjectivity and changing relations between humans and machines and especially between a particular person and a particular machine. He writes that "our ethical attention is shifted from ontology to epistemology, from object to subject, from 'what things really are' to how we look at things" (Coeckelbergh 2014, 66). Moral standing, he concludes, "is then not an abstract philosophical question but the practical question of how to relate and how to respond" (Coeckelbergh 2014, 66).

There is an important distinction to be drawn between the relationships found in Coeckelbergh's examples and hitchBOT's relationships with humans. Coeckelbergh's examples tend to be long term, personal relationships (e.g., eldercare robots), similar to Darling's social robots. During hitchBOT's travels, at least on initial contact, people did not have time to develop long-term relationships with it. Picking up a hitchhiker is more casual than establishing a friendship with, for example, a neighbor. Furthermore, there is no real opportunity for a person to develop a more long term or intimate relationship with hitchBOT – even to the level that one might have with a Roomba present in one's home. Yet, Coeckelbergh acknowledges that there could be multiple interactions and relations that contribute to the moral status of the robot, one that is defined in the "living phenomenology of daily experience ... the humans who 'meet' the robots, work with them, interact with them" (Coeckelbergh 2014, 67). In the next section, we look at applying Coeckelbergh's model to situations where the relationships are casual or even incidental such as when the relationships stem from meetings that happen on the street, in parks, and on the edges of our yards.

Before that, we consider the case when the unaccompanied encounter occurs when the owner or operator of that robot is also physically present. We draw on Paul de Laat's work to obtain some insight into how to consider these meetings (2016).

de Latt proposes that trust relationships can be established between the 'outside' human and the human/robot "coactive" team. He argues that the trust relationship between the two human agents can form a certain level of trust between the non-operator human and the robot. Arguing similarly, based on the moral standing that the two humans have, one might ascribe a certain moral standing to the owned robot. When it is clear that the robot is under the control of its human operator, it ought to have the moral standing that is typically ascribed to other property. For example, there seems to be a strong case that absent any other evidence or behavior, the robot ought not to be destroyed by the 'outside' human. Surely we would be taken aback if someone approached a stranger walking her dog and harmed the dog. The analogy is not perfect since a dog is a living thing and a robot is not; still, a robot with its owner is different in kind from a robot without an owner visible.

Again using an analogy, if we encounter a stray dog (with no associated human in evidence), we are likely to have a different reaction than to a similar dog that is the pet of a human who is physically present, especially when the dog is clearly under the control of the human. The pet/human "team" is quite a different experience for most passersby than an unaccompanied dog. Some humans may prefer a human-less dog if the dog is considered cute and friendly; the human may want to relate directly to the dog. But few people would prefer a human-less dog acting aggressively. Returning our attention to hitchBOT, hitchBOT's owners were not physically present when people encountered it. HitchBOT's simplicity and appearance were non-threatening and may have been essential for its success at establishing casual relationships with humans. Its non-threatening nature may also help explain people's reactions when hitchBOT was destroyed.

When a robot is accompanied by its owner/operator, the overarching relationship that directs the nature of the moral situation seems to be dependent on the nature of the relationship between the two people involved. Thus, we set this situation aside and in the next section we consider the moral relationship in casual encounters between people and robots when the owner/operator is not physically present and the person involved does not have a relationship with the owner through which the moral situation might be mediated. We will use the term "unaccompanied robot" to illustrate the importance of the absence of the robot's owners in a casual encounter.

17.3 Development Issues

For robot developers there are complex choices that surround the development of the interface. Darling notes that "people will ascribe agency to robots and treat them as social actors, particularly when they perceive them as lifelike entities rather than devices or tools" (Darling 2015, 1). In his analysis of Levinas and the principle of the other, Coeckelbergh seems to call for this approach to open the possibility that people perceive robots in new and different ways. He asserts that moral status goes beyond "I-you" relations as personified through the face of the robot. "How a particular robot appears to *me* (or how I construct it), and indeed how my concrete relation

with that robot is shaped, does depend on how we talk about robots (e.g. 'machines' versus 'companions'), on how *we* humans live together and live with robots, on the technological developments in *our* society, on *our* culture . . . " (Coeckelbergh 2014, 69). He asserts that if we have a relation with a robot, then the relation suggests that a moral standing exists and grows within the moral relation itself (Coeckelbergh 2014, 71). If we agree with Coeckelbergh that the "game of thinking about moral standing is itself dependent on the dynamics of concrete relations" (2014, 75), then we have to examine the role of those who develop robots and their interfaces. As robots become a part of our daily socio-technical environment, what kinds of ethical concerns should be addressed? Should we use a different paradigm in the development of social robots as opposed to non-social robots? If we focus on social robots like hitchBOT, how do we make ethical choices on the faces of the robots that seem to influence its relations with humans and its socio technical environment? By designing the face of the robot to "look like us", who exactly is the "us"? And, if robots resemble us, are we more likely to accept their actions without question, opening up possibilities of subversive activities such as privacy violations? How will the dynamic of the relate/respond relationship change? In the remainder of this section, we examine the development of non-social and fully-social robots, and the unaccompanied robots that are beginning to co-exist with us, addressing how their design influences their relations and hence their moral standing.

17.3.1 Non-social Robots

In the past, non-social robots were the norm in robotics. They were job-specific machines designed to perform particular tasks, often too dangerous for humans to undertake. Many were restricted to a particular geographic location because, for example, they were fastened to the floor. In those cases, human safety was ensured by preventing contact between humans and the robots. (Robots in a factory might be isolated from humans with cages around them.) The moral standing of the robot in a cage was made clear: it did not have any. Humans could not trust the robot and it (more accurately the robot's owners) could not trust that it would be treated appropriately by people. The sense of a separated *other* was physically expressed and enforced.

Another non-social robot is the robotic vacuum cleaner. While not nailed to the floor, it still is restricted to do its human owner's explicit and tightly prescribed bidding in someone's home. Any encounter with such a robot in the owner's home would be subject to the sort of analysis of de Laat's coactive team, even when the owner is not present.

There are more interesting cases of unaccompanied encounters with robots that are quickly becoming commonplace, and the unaccompanied nature distinguishes these encounters from the industrial robots and utilitarian robots just discussed.

Drones and driverless cars are not restricted to a particular geographic location and move further along the spectrum in the direction of fully-social robots. Additionally, while these devices do not currently come with the anthropomorphic framing that hitchBOT did, according to Darling it is not unreasonable for developers to expect people to anthropomorphize them. Proposed U.S. Federal Aviation Administration (FAA) rules for recreational drones say that the operator must maintain visual contact with the drone, suggesting at least an intuitive appreciation for the coactive team. Indeed, anecdotes of "stray drones," whose human operators are either remote or hidden, are part of the reason that the FAA is stepping in with regulations.

There is evidence that the supposed "perfect driver" of the driverless car initiates actions that are not anticipated by human drivers, and that this may lead to accidents, although they are less likely to be seriously injurious to humans (Naughton 2015). When a driverless car is empty, it is functioning as an unaccompanied robot. Any encounter a stranger might have with it is a casual encounter, in that the person was not seeking the contact (although the car itself may be moving with purpose).

Consider an encounter between an empty driverless vehicle and a person in a separate car. While there are legal and technical systems that can mediate this encounter, it is clear given the cases of human-to-human road rage that there is also a social element to any such human-to-driverless-car encounter. It seems that this is a reasonable testbed for Coeckelbergh's theory. This encounter is challenging for the theory, especially at night when one cannot easily discern whether the other car is a driverless car. How can we determine the moral status of the driverless car via relationships with humans when the humans in other cars are unlikely to realize that the car is driverless? If on encountering a car at night we assume that a human is driving that car, then we ascribe to the car's movements a human intentionality, and the driver's moral responsibility for the car's effects. If there is no human driver, then the observer is "relating" to the car under false pretenses, and any relationship is based on false assumptions. This seems difficult to resolve using only Coeckelbergh's theory.

We also foresee problems for Coeckelbergh's social-relational model in the following scenario: a driverless car senses the potential for an accident in its current situation, or in a situation that is likely to develop in the near future. Assume further that in this situation, the driverless car's accident-avoidance software has programmed in a moral ordering that places any humans and any human-driven cars higher than the empty driverless car. In this case, it is the developers' decisions, embedded in the software that are driving the driverless car's decisions. Using the social-relation theory seems insufficient to determine the driverless car's moral standing during the moments leading up to the potential crash. It is not the relationship between the driverless car and the humans that interact with it that determines whether its accident avoidance reflects ethically justifiable choices; it seems clear that the human developers' choices are either blameworthy or praiseworthy, regardless of any relationships between the machine and humans. Perhaps the social-relational model should be extended to the relationships between the humans who program the robot and the humans who are affected by the robot's actions; however, we don't see this emphasis in Coeckelbergh (2010) or

Coeckelbergh (2014). This case also seems to not be a good fit for de Laat's (2016) human/machine teams, since the developers are remote, in time and distance, from the driverless car during the moments of accident avoidance.

Another scenario is when the human is a pedestrian and the driverless vehicle is empty. In this case, the individual human seems to have lost footing to the robot car, at least in a physical sense. The amount of damage the car might do to the person exceeds that which the person might do to the car. It is clear that decisions of moral import are made by developers, not any humans directly relating to the car; for example, programming the car to hit a lamppost rather than a pedestrian in an accident scenario. As before, we see this as problematic for a social-relationship model, since neither the pedestrian nor the driverless car is likely to build a dynamic relationship; but it is likely that ethically significant effects may occur when a driverless car and pedestrians interact.

Finally consider an empty "driverless wheelchair" on its way to pick up its next passenger. This changes the power relationship from the one in the previous scenario. The amount of harm the chair can do to the human, while significant, is not as extreme as the potential damage from a car. Also, there is a more intimate human presence with a wheelchair. It represents a form of mobility for someone who is unable to walk. A wheelchair (clearly designed as a helping device for humans) may be less threatening than a car. These kinds of distinctions matter when people are experiencing machines and when people assign meaning to those experiences. In this case, the social-relationship model does seem promising; the human may evaluate the moral standing of an unaccompanied wheelchair (judged to be non-threatening and helpful) differently than the moral standing of an unaccompanied driverless car (judged to be large and potentially dangerous). However, the most important moral status relationship is between the human developers and the humans affected by the wheelchair, not the relationship between the wheelchair and the affected humans.

Notice that our motivating focus in this paper, hitchBOT, is designed differently than a driverless car or an automatic wheelchair. HitchBOT is not about transporting humans; hitchBOT is about being transported *by* humans. But neither is hitchBOT a robot that mimics human motions and emotions in any elaborate way. In the next subsection, we try to locate hitchBOT on a continuum based on how "socially" a robot behaves.

17.3.2 Fully-social Robots

It is useful to think of a continuum that stretches from robots that have little or no direct interactions with humans after deployment (like a welding robot) at one end, and robots that are designed to continuously interact with humans at the other end. We will call the constantly interacting robots "fully-social robots." In this subsection, we will make several stops along the continuum and consider how the

social-relation theory could be used to examine the moral status of robots with these different degrees of interaction with humans.

We do not think of hitchBOT as a fully-social robot because its interactions with humans are primitive. HitchBOT cannot gesture, walk, or engage in complex conversations. But hitchBOT was designed to be visually and behaviorally unthreatening and to interact with humans in a limited way. Therefore we locate hitchBOT somewhere between industrial robots and sophisticated, humanoid robots designed to mimic human's actions and speech while interacting with them.

HitchBOT's hitchhiking thumb, a culturally recognizable signal, elicits a response from those it meets. HitchBOT was given moral consideration by those who offered it a ride. Those who destroyed hitchBOT did not display an empathetic response to the machine, but it is difficult to say whether those humans anthropomorphized hitchBOT before they vandalized it. Regardless, the people who destroyed the machine clearly did not want hitchBOT to move through their neighborhood. They may have seen hitchBOT as a threat; at first it seems strange that someone would be threatened by such an outwardly passive device. But hitchBOT had surveillance and reporting capabilities. It also was a one-of-a-kind machine; people were unlikely to have seen a similar machine. It is not unusual for people to feel threatened by unusual things. The destruction of hitchBOT may well have been a primitive response to an outsider (if the humans anthropomorphized hitchBOT), or a hostile response to an unfamiliar machine (if the humans did not anthropomorphize hitchBOT).

HitchBOT's destruction also may have been a considered response: without permission, someone (remote and unknown) had injected a strange machine into someone else's environment. HitchBOT clearly had computing power, and it would have been reasonable to hypothesize that hitchBOT was a surveillance device. Although we may disagree with a violent response towards hitchBOT, it is not fanciful to suppose that someone encountering hitchBOT in their neighborhood might object to being surveilled by this machine. It may have appeared intrusive, as well as strange. Using the social-relation model, it may have been that those people in Philadelphia who encountered hitchBOT perceived it as a threat, and a relationship of animosity or fear developed, not a relationship of trust.

The eventual destruction of hitchBOT is cautionary for applying the social-relational model to help determine the moral status of an entity. Social relations are not always positive; some social relations are intensely negative. If we are to look for moral status based on dynamic social relations, we should be prepared to deal with negative as well as positive relations. And we should take into account that a machine with negative relations with humans may be regarded by some with hostility, as an "other" with little or no moral status. Hostile reactions to hitchBOT (which did not have to be violent) could be understood, and even rationally defended, based on a social-relational model.

Several design decisions by hitchBOT's developers may be ethically problematic. In the design of the robot and its human interface, and in the design of the website devoted to hitchBOT, the developers encouraged a pretense that hitchBOT had human emotions and motivations. We expect that hitchBOT's developers

presented this fictional sophistication as playful, not deceptive. However benign, this fiction does have some ethical problems; anthropomorphizing a machine can have serious consequences, especially when the attitude of society towards robots is significantly affected, encouraging the general perception of robots to become unrealistic. Returning to the social-relational theory, basing a human-robot relationship on playful exaggerations is problematic.

Many robot pets are more socially adept than hitchBOT. They are capable of sophisticated movements and sound, and can learn behaviors in order to be more responsive to humans. This example is cited by Coeckelbergh (2014) and Turkle (2011) when they analyze socially relevant interactions between machines and humans that lead to a person becoming emotionally attached to a robot pet. However those analyses do not consider the casual encounter another person might have with an unaccompanied robot pet. The developers of such robots have a clear role to play in the social relationship that manifests itself. That role goes beyond the framing that occurs due to the typically attractive animal-like design. There are design issues that have to do with how the robot interacts with other people when the owner is not present. Decisions about how active or passive the robot is in such encounters will influence the moral status a person ascribes to the robot. Yet the developer must take care with the assumptions made about people regarding their attitudes toward casual encounters with animals.

Our final stop along the continuum is an encounter with an unaccompanied sophisticated humanoid looking robot, a robot with Internet connectivity. Many of the questions raised above with other robots become more complex when the robot looks more humanoid, because issues of anthropomorphic framing described by Darling (2015) increasingly come into play during social interactions. These are complicated because of things that we have described previously, including the remote (social) relationship with the developer.

There is one situation that we have not previously considered: the relationships that arise when a hacker gains a certain level of control over any of the robots we have examined. Before a hacker invades, there are three relationships of interest in an analysis: the developers and their robot, the robot and the human it encounters, and the indirect relationship between the human encountered and the developers. When a hacker becomes part of the situation, there are three new relationships, each with its own nuances. Two of these new relationships are strictly human: the hacker and the developers, mediated through the robot; and the hacker and the encountered human, again mediated through the robot. The third new relationship is the robot and the hacker. We will not go into detail about these relationships, but we will mention that the hacker/robot relationship suggests that the robot can be classified as a moral patient, a victim of the hacker, rather than as a moral agent.

The closer we get to interfaces for robots that approximate human interactions, the easier it is for people to respond and empathize. Developers who design robots with rigid, clearly defined social roles have easier choices to make. Developers who design robots that coexist with humans and have no single prescribed function, will have more (and more complicated) ethical choices to make. Developers will need to consider the implications of an unaccompanied encounter. As robots begin to pass as

human (even if only from a distance), there will be many interesting legal and ethical issues that will quickly gain importance. We expect that using the social-relational model will be useful (though not a panacea) for working through those issues.

17.4 Universalization

We briefly consider "universalization" as applied to unaccompanied robots. That is, it may be fine to have a few robots, like hitchBOT, that garner our attention. While it is rare to encounter unaccompanied robots, it seems tractable to work through ethical issues about covert surveillance, resource allocation, and benign deceptions. But what if unaccompanied robots become commonplace? How does that impact the tractability of the ethical issues? Is there a tipping point after which the sheer number of unaccompanied robots around us becomes an ethical issue? It seems clear that resource allocation would be an obvious problem as the number of robots became large. But are there other problems that are candidates for Coeckelbergh's social-relational analysis?

First let us consider the case of driverless vehicles – many of them on the road, some carrying people, others not, and just a few people driving their own cars. Here Coeckelbergh's approach may offer some interesting insights. If it does play out empirically that driverless cars lead to fewer and less severe accidents, then it seems that there is a case for the driverless car to be held in higher moral regard than vehicles with a human driver. When a vehicle with a human driver is demonstrably a bigger threat to human and driverless-vehicle flourishing than the driverless vehicle, our moral perspective shifts. The driverless vehicle, as a largely technical system, then will be a better fit for the socio-technical system of roads and driving rules, at least with respect to safety. One consideration is that the sociotechnical system of roads and travel will be less social and more technical if driverless cars come to dominate the system. However, the decision to largely replace human drivers with robot drivers would itself be an intensely social decision, with significant political, economic, and legal ramifications.

Another extreme is when we find our parks, sidewalks, and boulevards crowded with humanoid looking robots that behave largely like we expect people to behave. At least one interpretation of Coeckelbergh's approach suggests that we should give up any sort of discomfort we might find with granting a robot, even an unaccompanied one, a clean slate in order to establish a relationship with the robots we encounter. But, if the number of robots we encounter on this rather personal level approaches the number of humans we encounter, it may be detrimental to society. So many robots competing for our attention (and perhaps for our affections) could have serious consequences for humanity. The social relationships between humans could suffer; indeed, it is a common fear that mobile phones, computer games, and various electronic entertainments have already degraded human-to-human relationships. Surely this concern will become increasingly worrisome when humanoid robots are ubiquitous. And just as surely, if a degradation occurs, it will have serious ethical ramifications for the thriving of humans.

17.5 Conclusions

We have taken Coeckelbergh's theoretical work and Darling's applied work and considered both in the context of human/robot social relationships. We have seen that different sorts of robots lead to different sorts of social relationships and that these differences are ethically important.

We see hitchBOT as a particularly interesting example. We focused on the relationships that hitchBOT's developers designed hitchBOT to encourage. For many people who encountered hitchBOT, they seemed to form a relationship that was positive, though temporary. Eventually, someone who encountered hitchBOT physically destroyed the robot. We speculated on why this might have occurred and examined some problematic aspects of hitchBOT as an unaccompanied robot and the hitchBOT website that might have been relevant to attitudes towards the robot. We were particularly interested in issues of benign deception, potential covert surveillance through unaccompanied robots, and the proliferation of robotic others.

We are convinced that the work of Coeckelbergh, Darling and de Laat can be helpful in analyzing human/robot ethical issues, though applying their insights does not remove all potential difficulties with those relationships or with our ability to clearly understand their ethical importance. Extensions of their work that consider relationships between people and robots when the owner or operator of the robot is not present would be valuable. Focusing on the social relationships between humans and robots does *not* remove responsibility from a robot developer or owner for that robot's behavior. On the contrary, knowing that humans may form significant relationships with these machines *increases* developers' ethical responsibilities to ensure (as much as is practical) that those relationships will be positive. The potential for the robot to encounter other people when its owner is not present further complicates the ethical calculus for the developer. The possibility of hackers taking control of robots highlights the responsibility of developers for ensuring adequate security for unaccompanied, sociable robots. And we note that hackers greatly complicate both the situation of human/robot relationships and any attempt to analyze the ethical ramifications of these relationships.

References

Coeckelbergh, Mark. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12 (3): 209–221.

———. 2014. The moral standing of machines: Towards a relational and non-Cartesian moral hermeneutics. *Philosophy & Technology* 27 (1): 61–77.

Darling, Kate. 2015. 'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy. *Proceedings of WE Robot Conference on Robotics, Law & Policy 2015*.

de Laat, Paul B. 2016. Trusting the (ro)botic other: By assumption. *ACM SIGCAS Computers and Society* 45 (3): 255–260.

Grodzinsky, Frances S., Keith W. Miller, and Marty J. Wolf. 2011. Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Ethics and Information Technology* 13 (1): 17–27.

———. 2015. Developing automated deceptions and the impact on trust. *Ethics and Information Technology* 28 (1): 91–105.

hitchBOT: A robot exploring the world. 2015. www.hitchbot.me/about. Accessed 28 Dec 2015.

Naughton, K. 2015. Humans are slamming into driverless cars and exposing a key flaw. *BloombergNews*. www.bloomberg.com/news/articles/2015-12-18/humans-are-slamming-into-driverless-cars-and-exposing-a-key-flaw. Accessed 8 Jan 2016.

Turkle, Sherry. 2011. *Alone together*. New York: Basic Books.

VanderMaas, Johanna. 2015. *hitchBOT USA tour comes to an early end in Philadelphia*. cdn1.hitchbot.me/wp-content/uploads/2015/08/hitchBOT-USA-Trip-End-Press-Release-FINAL.pdf. Accessed 28 Dec 2015.

Chapter 18
Against Human Exceptionalism: Environmental Ethics and the Machine Question

Migle Laukyte

Abstract This paper offers an approach for addressing the question of how to deal with artificially intelligent entities, such as robots, mindclones, androids, or any other entity having human features. I argue that to this end we can draw on the insights offered by environmental ethics, suggesting that artificially intelligent entities ought to be considered not as entities that are extraneous to the human social environment, but as forming an integral part of that environment. In making this argument I take a radical strand of environmental ethics, namely, Deep Ecology, which sees all entities as existing in an inter-relational environment: I thus reject any "firm ontological divide in the field of existence" (Fox W, Deep ecology: A new philosophy of our time? In: Light A, Rolston III H (eds) Environmental ethics: An anthologyBlackwell, Oxford, 252–261, 2003) and on that basis I introduce principles of biospherical egalitarianism, diversity, and symbiosis (Naess A, Inquiry 16(1):95–100, 1973). Environmental ethics makes the case that humans ought to "include within the realms of recognition and respect the previously marginalized and oppressed" ((Gottlieb RS, Introduction. In: Merchant C (ed) Ecology. Humanity Books, Amherst, pp ix–xi, 1999)). I thus consider (a) whether artificially intelligent entities can be described along these lines, as somehow "marginalized" or "oppressed," (b) whether there are grounds for extending to them the kind of recognition that such a description would seem to call for, and (c) whether Deep Ecology could reasonably be interpreted in such a way that it apply to artificially intelligent entities.

Keywords Moral responsibility · Environmental ethics · Deep ecology · Artificial intelligence · Artificial agency

M. Laukyte (✉)
Department of Private Law, Universidad Carlos III de Madrid, Madrid, Spain
e-mail: migle.laukyte@uc3m.es

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_18

18.1 Introduction

Time and again we have raised the question of how in the future we ought to treat artificially intelligent entities, such as robots, mindclones, androids, bemans, or any other entity having intelligence, autonomy, or other features that would make it similar to a human being.[1] Furthermore, with human enhancement, and with the prospect of technologies like those that try to build robots with a biological brain grown in an incubator or to upload a human brain onto a computer (Kurzweil 2006; Rothblatt 2014), it is no longer ontologically clear what it is to be human or how we should draw the line between human and nonhuman, and where we should place transhumans, namely, individuals who "transcend human biological inheritance, modifying their DNA, their bodies, or the substrate for their minds" (Rothblatt 2014, 307).

These technological scenarios confront us with the ethical problem of inclusion and exclusion: Are the new entities worthy of consideration as moral beings? And, if so, on what basis? Depending on the way we answer these questions, we will come out with different ways of treating these new entities, thus fundamentally shaping the social environment in which we are going to live in the future and which we are going to pass on to the future generations.[2]

This paper offers an approach to the question of how to deal with artificially intelligent entities: I propose that we draw on the insights offered by environmental ethics, suggesting that artificially intelligent entities ought to be considered not as entities extraneous to our social environment, but as forming an integral part of that environment. The argument I will be unpacking builds on the radical strand of environmental ethics known as Deep Ecology,[3] whose underlying premise is that all entities exist in an inter-relational environment: Deep Ecology thus rejects any "firm ontological divide in the field of existence" (Fox 2003, 255), and on that basis it introduces principles of biospherical egalitarianism, diversity, and symbiosis (Naess 1973). Environmental ethics makes the case that we humans ought to "include within the realms of recognition and respect the previously marginalized and oppressed" (Gottlieb 1999, ix), so in this paper I consider whether (a) artificially intelligent entities can be described along these lines, as somehow "marginalized" or "oppressed"; (b) whether there are grounds for extending to them the kind of

[1] A mindclone is a cyberversion of a human being, with a human mind uploaded on a digital support, whereas a beman is not a replication of human mind but an entity that is cyberconscious on its own account. On mindclones, bemans, and other possibilities offered by artificial intelligence, see Rothblatt (2014). In the interest of clarity, I will use the term *human* to refer to human beings, *nonhuman* to refer all other living and nonliving entities (animals, mountains, rocks, machines), and *artificial* to refer to artificially intelligent entities and other artificial forms of life.

[2] On the moral treatment of artificial agents and its justification on different grounds, such as rationality, interactivity, and autonomy, see Floridi and Sanders (2004), Tavani (2011) and Coeckelbergh (2009 and 2010).

[3] Other radical theories are Social Ecology, Political Ecology, and Ecofeminism (Keulartz 1995).

recognition that such a description would seem to call for; and (c) whether the ideas of Deep Ecology can be applied to artificially intelligent entities.

The discussion is organized as follows: In Sect. 17.1, some of the key notions, related to artificially intelligent entities, environmental ethics, and Deep Ecology, are explained. In Sect. 17.2, the focus is on why and how the ideas of Deep Ecology could apply to artificial intelligence, focusing in particular on some of the eight principles of Deep Ecology: the argument is that these principles are applicable not only to biological entities and the biosphere in general, but also to artificially intelligent entities. In Sec. 17.3, the focus shifts to the main difficulty with the idea of bringing Deep Ecology to bear on artificially intelligent entities. This is the idea that nature—or the environment at large—is a breathing and evolving organism made of living sentient entities (such as animals, fish, and plants) and as such is thus worthy of moral consideration. The difficulty is that this description—namely, being alive or sentient—is usually not attributed to artificially intelligent entities. This critical point is addressed by offering a way out of the impasse, arguing that being alive and sentient are not essential requirements for moral consideration, while pointing out alternative approaches that have been developed in that regard, so much so that even Deep Ecologists themselves as well as many environmental ethicists agree that landscapes and mountains, for example, are also worthy of moral consideration. Having addressed those issues, the paper finishes with a few closing remarks.

18.2 Some Notes on Terminology: Who's Who?

Before taking up the arguments for and against extending Deep Ecology from the natural environment to artificially intelligent entities—from the natural world to the artificial world, thus providing the concept of the environment with a new and more inclusive meaning we need to make some clarifications about the terminology used in this paper.

I begin with the idea of an artificially intelligent entity. As suggested earlier, artificially intelligent entities are any kind of entity having an artificially built intelligence and other features associated with intelligence, such as autonomy and the ability to make reasoned decisions. This artificial intelligence I regard as similar to human intelligence: It may outsmart human beings in some respects (Bostrom 2014), while falling short in others. The point is not to rank different forms of intelligence on any scale of excellence: It is rather to determine whether they have the kinds of features that would trigger the question of moral consideration. This means that it does not matter how this intelligence is achieved: It can be via human whole-brain emulation or by uploading a human brain onto a digital device (Rothblatt 2014)—or indeed any embodiment of artificial intelligence, be it digital, virtual, or physical—so long as it resembles human intelligence. This importantly means that I regard it as essential to artificially intelligent entities that they have social or interactive capacities, and that in light of that behaviour we can ascribe some kind of emotion or other to them. The artificially intelligent entities taken

into account here can thus be described as embodying a general-purpose artificial intelligence, an intelligence that is not related to any particular task but applies across different environments and contexts and to a range of different problems (just like human intelligence).

A problem comes up in regard to the sentience of such entities: are they sentient or nonsentient? This is the criterion by which we usually determines whether we have a duty of ethical consideration: If the entity is sentient, we owe some consideration; if it is not, we can exploit it in any way that we think is beneficial to us (this I will call the anthropocentric view). There are many reasons why this way of thinking is wrong, but I will focus on two of them. For one thing, Deep Ecologists already acknowledge that nonliving entities, such as mountains, are inherently worthy of moral consideration, so it is beside the point whether or not artificially intelligent entities are sentient. And, for another, by developing artificially intelligent entities, we might also develop a different and new kind of sentience that will challenge our idea of what sentience and nonsentience are.

Let us turn now to environmental ethics: This is the branch of ethics that focuses on the interaction between humans and "nonhuman nature within the context of ecological systems" (Keller 2010, 3). It branches into several subareas, but what links them all together is the juxtaposition of, and competition between, two values that are attributed to nature, namely, its instrumental value (nature as a means to an end) and its intrinsic value (nature as an end in itself).[4]

We can now consider Deep Ecology.[5] This is a field of environmental ethics that departs from mainstream environmentalism by moving away from the previously mentioned anthropocentric view on which nature is worthy of protection only insofar as that is instrumental to human welfare. Deep Ecology, by contrast, envisions a deeper way of dealing with environmental issues, not only from a philosophical perspective but also from a political one. It does so by looking at nature as valuable in itself, regardless of whether it is useful to human beings: It thus assigns intrinsic value to ecosystems (Baard 2015). This view has been termed biospheric egalitarianism. And the reason why it describes itself as deep is that, in reframing our understanding of nature, it calls on us to fundamentally change our way of relating to it, not as a means to an end but as an end it itself. This can be achieved by "engaging in a process of ever-widening identification with others" (Keulartz 1995, 118), an identification which is not be limited to other human beings but extends to the entire biosphere, and which would therefore be impossible without "a more sensitive openness to ourselves and nonhuman life around us" (Devall and Sessions 1985, 65).[6]

[4]Environmental ethics quite often deals with the juxtaposition between anthropomorphism and nonanthropomorphism, holism, individualism, and other ideas (Keulartz 1995).

[5]The seminal study on Deep Ecology is Naess (1973), and the view has since been developed in numerous works. For an overview, see Naess (1986, [1989] 2001, 2005, 2008) and Keller (2008, 206–11).

[6]In making this identification, however, Deep Ecology does not argue that all beings enjoy the same moral standing. Naess himself concedes that a ranking of beings is inevitable, pointing out that this

Deep Ecology is based on eight principles premised on the idea that humans no longer form the centre of discourse. This does not amount to removing the human being from the spectrum of moral consideration, but it does mean that since the human being is deeply intertwined with nature, the two components of this relation—namely, humans and nature—are to be regarded as forming a whole rather than as separate entities.

Let us see, then, what these eight principles of Deep Ecology are:

1. The well-being and flourishing of human and nonhuman life on Earth are valuable in themselves, from which it follows that the value of nonhuman life-forms is not a function of its usefulness to humans.
2. The flourishing of human and nonhuman life is dependent on the richness and diversity of life forms, and this diversity is itself inherently valuable.
3. The inherent value of the richness and diversity human and nonhuman life means that humans do not have a right to reduce such richness and diversity, except to satisfy vital needs.
4. Human life and cultures can flourish even with a substantially smaller human population.
5. Human interference with the nonhuman world is excessive.
6. We must therefore make structural economic, technological, and policy changes.
7. The underlying change will have to be ideological: a change in attitude that consists in appreciating the quality of life itself rather than aiming for an increasingly higher standard of living as measured by economic growth.
8. The foregoing principles entail a duty to join together in an effort to implement the necessary changes (Naess 1986, 2; 2008, 111–12).

But before explaining how these principles could be applied to artificially intelligent entities (in Sect. 17.2), we still have a more fundamental question to address, namely, why choose environmental ethics, and Deep Ecology in particular, as a basis for reasoning about artificially intelligent entities?

Let us first consider three main reasons for framing the discussion on the basis of environmental ethics. The first reason is that we do not yet have any sufficiently broad ethics for artificial intelligence: Azimov's three laws of robotics cannot help us solve this problem, so we need a more solid ground on which to build an ethical approach to artificial intelligence. And the second reason is that we want to avoid the errors we made in the past in framing an ethical approach to nonhuman entities. This suggests looking for moral guidance outside the realm of specific disciplines, such as the philosophy of technology or the philosophy of artificial intelligence. Environmental ethics makes it possible to give broad scope to the question of the moral consideration of nature and the environment, and to do so in such a way as to address the paradigmatic shift we confront in the human approach to the ecosystem, in that the "biosphere [. . .] has become a human trust and has something of a moral

also entails a raking of duties: The duties we owe to fellow human beings are higher than those we owe to other beings, such as mice. This is an easy choice—humans versus mice—but there are choices that are neither clear nor easy, especially when it comes to ranking different species. This is why Naess (2005) describes ranking a complex process, and not straightforwardly moral.

claim on us" (Jonas 1984, 8). There is, finally, a third reason why environmental ethics seem promising as an approach to the ethics of artificial intelligence: This relatively young and dynamic field of moral inquiry takes in different ideas from the other moral theories,[7] and in so doing it offers some insights that can be helpful in dealing with nonhuman otherness.

The idea of drawing on environmental ethics to address the moral problem of artificial intelligence is not new. Gunkel, for example, points out that environmental philosophy, animal rights philosophy, and the machine question all seek "to think outside the restrictions of anthropocentric privilege and human exceptionalism" and consequently "to dissolve the kind of human centric view of the universe that is being broken open by what we can say is a Copernican Revolution" (quoted in Kellogg 2014). He turns to environmental philosophy because in it he finds "a thinking of otherness that is no longer tied to either human centrism or biocentrism" (ibid.).[8]

But why Deep Ecology in particular? I will point out four reasons. First, seeing the human-centric (or anthropocentric) approach to environmental ethics as problematic, Deep Ecology takes an ecocentric approach: Instead of placing the human being at the centre of the discussion, it places humans next to other (biological) entities. Accordingly, Deep Ecology rejects the position that regards "humans as isolated and fundamentally separate from the rest of Nature, as superior to, and in charge of, the rest of creation" (Devall and Sessions 1985, 65). This is a good starting point, because it enables us to address the issue of artificially intelligent entities without narrowly preselecting humans as the main lens through which to understand what is worthy of moral consideration.

Second, Deep Ecology does not confine itself to strictly philosophical inquiry but advocates a wider and more profound social change: It takes us from a purely theoretical discussion to a more practical level, asking us to consider the need for institutional and political change, and that is exactly what may be needed in dealing with artificially intelligent entities (Bostrom 2014; Rothblatt 2014; Kurzweil 2006).

Third, Deep Ecology proceeds not from normative prescriptions but from principles. Unlike prescriptions, principles are broad and flexible, making it possible to interpret and shape them in ways that will meet the demands of a discussion on artificially intelligent entities. And, as we will see in Sect. 17.2, a useful link can be established between Deep Ecology and artificial intelligence (see Coeckelbergh 2010).[9]

[7]Thus, for example, environmental ethics introduced the question of justice in the debate on environmental problems. On this development, see Armstrong (2012). The problem with traditional theories is that they are all anthropocentric and no longer adequate to deal with current problems (Troster 2008, 392), but that need not be the case, considering that Deep Ecology draws inspiration from well-established theories like those of Spinosa, Heidegger, and Whitehead (Keulartz 1995).

[8]In this connection, see Coeckelberg (2010) and Rothblatt (2014), drawing parallels between robot ethics and animal ethics in a way that brings environmental ethics into the picture.

[9]The bearing that Deep Ecology has on artificial intelligence and computer ethics is also briefly mentioned in Floridi and Sanders (2001).

Finally, the fourth reason is that, unlike anthropocentrism, with its short-term vision of environmental problems, Deep Ecology takes the long view (Baard 2015): this is precisely what we need if we want to have any kind of discussion about the future of artificial intelligence and its place in our human and natural environment. If we are to work toward any kind of fruitful coexistence of human, natural, and artificial life-forms, we need to extend our view over the long stretch. Indeed, short-term thinking would make the discussion irrelevant from the start, considering that artificially intelligent entities of the kinds that would make these problems real have yet to be created.

In the following Sects. 17.2 and 17.3, I will introduce arguments for and against applying Deep Ecology to artificially intelligent entities, exploring the reasons why the principles of Deep Ecology could extend to artificially intelligent entities, and considering how the arguments against such an extension could be defeated.

18.3 Deep Ecology as an Approach to the Problem of Artificially Intelligent Entities

Let us consider the arguments in favour of applying the principles of Deep Ecology to some of the challenges that artificially intelligent entities may give rise to in the future. I will argue that the insights Deep Ecology offers in dealing with the environment and the moral status of nature can also shed light on the question of the moral, social, and political implications of artificial intelligence. I begin by pointing out that, while the object of discussion may different (the environment and nature as against artificial intelligence), the problem is the same (exploitation) and so is the decision-making entity (the human being). I elaborate on this point by taking the eight principles of Deep Ecology and applying them to artificial intelligence so as to see whether these principles are applicable to something more inclusive than nature, the environment, and the ecosystem.

As we saw, the first principle invites us to consider human and nonhuman life—or, as Jonas (1984, 8) puts it, "extrahuman" life—as intrinsically valuable, regardless of its contribution to human welfare. I submit that this principle can be extended to artificially intelligent entities because they, too, can be seen as a form of nonhuman life. Note that *life*, in the term *human and nonhuman life*, is understood by Deep Ecology to include rivers and landscapes: these are "nonliving" (Devall and Sessions 1985), and so also nonhuman, forms of life. And if rivers and landscapes are considered in this way—as nonhuman forms of life—so can artificially intelligent entities, as artificial forms of life.[10] So the question here is not What is valuable to

[10]Here is what one commentator has written on the prospect of artificial life: "Many agree it is only a matter of time before artificial life creates machines that are alive, are intelligent, reproduce their own kind, have their own purposes, set their own goals, and evolve autonomously. These machines will be as much a part of the natural world as features in the landscape or existing forms of life,

human life? but How can human and nonhuman life, including artificial or synthetic life, be made compatible, and indeed coherent, weaving into a single fabric the value that can be recognized as intrinsic to all?

The same applies to the second principle. This principle recognizes the richness and diversity of life-forms, and artificially intelligent entities can be counted as a life-form, however much artificial or synthetic. These life-forms are not yet known to us, but on a Deep Ecology approach they can be regarded as valuable in themselves, just like other life-forms.

On the third principle, human beings do not have a right to reduce the richness and diversity of life-forms except to satisfy their vital (existential) needs, and for no other reason, and on the fourth principle human society can flourish consistently with a smaller human population: if we extend these two principles to the moral question of our treatment of artificially intelligent entities, we can see that there seem to be no vital needs in virtue of which to justify reducing the diversity of an environment inclusive of these entities, nor do these entities seem to have much influence on the growth of the human population: If artificially intelligent entities contribute to the richness and diversity of life-forms, and if they bear little relation to population growth, they are protected under these two principles.

The fifth principle asks us to reduce human interference in the nonhuman world. This principle raises something of a paradox because, if on the one hand such human interference is in large part responsible for the environmental problem we face today, on the other hand humans need to keep interfering in the nonhuman world so as to deal with and solve that very problem. The paradox is solved, then, by looking at the "nature and extent of such inference" (Devall and Sessions 1985, 72): Not all interference is of the same kind or equally extensive. This principle is more difficult to apply to artificially intelligent entities than the other principles of Deep Ecology because, on the one hand, artificially intelligent entities are artificially created, and so anything they do is ipso facto artificial, but on the other hand, they are so inextricably bound up with their human makers, and the interaction is so close, that it is difficult to draw a neat line of separation.[11] I would therefore count this as the most problematic principle of Deep Ecology. And the problem is compounded by the fact that the boundary between beneficial and harmful human interference is blurred, such that, when dealing with artificial intelligence, we probably need to make case-by-case judgments.

The sixth principle calls for a structural change in policy, moving away from a laissez faire model of self-regulating production, consumption, and growth that does not concern itself with the problem of externalities—i.e., the social and environmental costs of free-market capitalism—toward an environmentally sustainable model

and their evolution will affect the course of existing forms of life. [...] machines might play an unprecedented role in the next major evolutionary transition, and the challenge here is to predict and explain this role. Machines may well be the central players in the transition, as will be the case if autonomously evolving machines get established in the natural world" (Bedau et al. 2000, 373).

[11]The interaction between humans and artificially intelligent entities will be developed further in the next section.

that does take those costs into account. In this respect, too, artificially intelligent entities can be seen to play a dual role. For on the one hand, as a product of this free-market model, they are part of the problem, but on the other, as a technology, they can also be part of the solution, contributing to the paradigm shift toward sustainability.[12]

Closely bound up with the sixth principle is the seventh, positing an inherent quality of life that cannot be reduced to the standard of living, in which the attempt is to measure the quality of life by the amount of goods and services produced in the economy, that is, by the total output of the economy, or gross domestic product, a measure plagued by the problem that it counts any economic transaction as growth regardless of whether it is sustainable or unsustainable. On the seventh principle, the idea is that while the quality of life may be dependent on material quality (as measured by access to goods and services), it cannot be *equated* with this measure, nor can it be severed from the problem of the whole—the problem of what the acquisition of material quality entails for the good of the planet as an interconnected whole. We can see that this principle clearly applies to artificial intelligence: even if Naess warns us against this neophilia (Baard 2015), the technology can be used to improve the quality of life—by relieving humans of the burden of carrying out tasks that do not seem to have any inherent value—and it can do so in an environmentally sustainable way.

The eighth principle calls on us to implement the first seven. Naess ([1989] 2001, 26, 45) observes that this would requires "a substantial reorientation of our whole civilization," with "new criteria for progress, efficiency, and rational action," and "new social forms for co-existence." This rethinking of society and civilization is clearly open-ended and open to interpretation, and one can expect a good deal of disagreement over the practical details, but there is no doubt that in the solution we can fit the idea of our coexistence with artificially intelligent entities.

What we can appreciate from this rundown is that there is no principled reason why we should be prevented from applying the principles of Deep Ecology to the question of artificially intelligent entities and our treatment of them. On the contrary, these principles can be useful from the outset in framing the issue of artificially intelligent entities in a constructive way. The issue is not so much about these entities themselves as it is about us, how we ought to interact with them, and the place we should find for them. If there is an overarching principle that captures the whole of Deep Ecology, it is that we live in a holistic system of interdependent components that are valuable in themselves, and whose interaction is essential to the life of the system itself: although the standing paradigm of social organization based on the idea of the market economy as a self-regulating system seems consistent with that overarching principle, we have learned from experience that a literal application

[12] As a societal model on which to base our interaction with nature, sustainability is also advocated in Kortetmäki 2016. Even if sustainability is consistent with the policy changes called for under the sixth principle, Naess was critical of the idea as such, arguing that it is anthropocentric and therefore out of keeping with the holism by which Deep Ecology is underpinned (Baard 2015).

of this idea comes with costs that are both social and environmental. Deep Ecology offers a way to reconsider that idea in such a way as not to repeat the mistakes of the past, and artificial intelligence can certainly be part of that solution.

That, in a snapshot, is Deep Ecology in connection with the problem of artificially intelligent entities. But there are a couple of important arguments that work against this approach. Let us therefore see what these arguments are and how we might respond.

18.4 A Critique of Deep Ecology

In this section we will consider two arguments against the idea of drawing on Deep Ecology as an approach to the issue of how we ought to relate to artificially intelligent entities. The first is the argument that Deep Ecology is ideally suited to dealing with conscious, sentient, or living organisms as part of our social and legal environment, and that these are not characteristics we can use to describe artificially intelligent entities.

There are many commentators who have responded to this objection. Thus, Rothblatt (2014) draws a parallel between consciousness and the cyberconsciousness ascribable to our mindclones, arguing that different entities have different forms of consciousness, and that these differences are irrelevant to whether artificially intelligent entities can be regarded as worthy of moral consideration, and hence whether there are reasons for bringing them under the protection of the law. Developments in synthetic (artificial) life applications suggest that the same reasoning applies when noting the property of being alive and sentient as obstacles to moral consideration. Nor does any other biological sort of property seem to keep us from recognizing artificially intelligent entities as having a moral status (Floridi and Sanders 2004).

The argument against such moral recognition seems to take an approach that consists in checking off a list of properties acting as necessary conditions to be met, but some alternative approaches have been proposed. One example is Coeckelbergh (2010), suggesting that we attack the problem by focusing on the social relations between humans and robots, without having to look at the properties ascribable to robots or their ontological features. This relational approach suggests that we would treat industrial robots in one way and domestic robot assistants in another.

The same approach is proposed by Gunkel (Kellogg 2014), who points out the increasing social interactivity of robots. Nor is this approach confined to philosophical inquiry: Engineers, too, have recognized how important it is to take the social aspects into account in dealing with human interaction with artificially intelligent entities. Thus, Farshchi (2016) argues that we should switch from building human-machine interfaces to creating machine-human interfaces. Where does the difference lie? A human-machine interface is based on natural-language processing and speech recognition, while in building a machine-human interface we are focused on understanding people and their emotional states. Such emotion-reading interfaces would hugely contribute to building social relations between

machines and human beings. Examples are JIBO, the world's first social home robot, which "communicates and expresses using natural, social and emotive cues"[13]; Pepper, whose "number one quality is his ability to perceive emotions" and adapt accordingly via an Emotion Engine[14]; and CoBot robots, based on symbiotic human-machine interaction.[15] All these examples show that research in robotics is appreciating the crucial role of machine social skills, which are indispensable if we are to achieve a deeper interaction with greater empathy between human beings and machines.

The relational approach—as opposed to the argument that moral consideration necessarily requires sentience and biological life—also finds support in the work of ecologists such as Kortetmäki (2016, 92), who takes the example of the lakewater pollution caused by the Talvivaara nickel mine in Finland: In making the case that lakes are worthy of moral consideration, she does not stress that they are valuable as ecosystems on which we depend, or that the environmental damage done to them is detrimental to our health or kills a variety of living organisms, but rather argues that the issue is about "the lakes themselves as places to which the people have special relations." The same special relationship humans establish not only with places but also with the artificially constructed world, and even more so with artificially intelligent entities.

This relational approach to artificial intelligence—offering an alternative to the property-based approach—might seem inconsistent with Deep Ecology on account of the eight principles, which seem to work as a checklist. But that is not what we should take away from Deep Ecology. Indeed, its central insights on the question of whether other entities (natural or artificial) ought to be recognized as having a moral status revolve around the appreciation that we share the same interactive environment with them.

The second of the two previously mentioned arguments against the idea of applying Deep Ecology to artificial intelligence raises a problem of coherence. The argument proceeds from the fundamental distinction between nature and artificially intelligent entities: Nature is not a human creation, while artificially intelligent entities are. Ergo: If it is wrong for humans to exploit nature, why should it be right do so with something (or someone) they created themselves and which (or who) would not exist without the human beings that developed them in the first place?

The flaw in this argument lies in its factual premise, in that a large chunk of nature *is* in fact created by human beings. Let us consider domestic animals, like dogs. Many dog breeds are created by human beings, and if it weren't for such human breeding, those breeds would not exist. The same applies to different kinds of other animals and plants developed using different techniques (such as genetic engineering and plant breeding). So, on this reasoning, we should draw a distinction among nonhuman life forms that are developed by human beings and nonhuman

[13]More information about Jibo is available at www.jibo.com

[14]More information available at https://www.aldebaran.com/en/cool-robots/pepper

[15]More information available at http://www.cs.cmu.edu/~coral/projects/cobot/

life forms whose development is not owed to human intervention, and we should therefore preserve the latter and ignore the former. A moment's thought, however, should suggest that it may not be a good idea to extract moral consequences from such a distinction: The "authorship" of some species is not a license to treat these species however we like. This applies to animals and plants, and it should also apply to artificially intelligent entities.

Furthermore, if we want to discuss the role that humans play in the nonhuman world, we should frame the discussion in terms not of authorship but of stewardship, which implies a duty of care and responsibility to nonhumans: This is what Jonas (1984) argues as concerns nature, and what Floridi and Sanders (2004) argue as concerns artificial agents. And although Deep Ecologists do not like the idea of stewardship (Devall and Sessions 1985), the idea may well serve as another starting point for a discussion of what it is to relate to that which surrounds us.

18.5 Conclusions

In this paper we considered the question of how we ought to relate to artificially intelligent entities as entities forming part of our natural and constructed environment. It was suggested that that one solution may come from Deep Ecology, a theory that in the description of its founder, Arne Naess, "asks deeper questions: we ask why and how, where others do not" (Devall and Sessions 1985, 74). This is why I believe that Deep Ecology can provide an interesting lens through which to discuss the moral standing of artificially intelligent entities: we see these entities as things, and we seldom ask the deeper questions that go beyond the anthropocentric conception, from which comes a spectrum of stances ranging from unfiltered consumerism to the antagonistic "we-against-them" mindset.[16]

The underlying idea of this paper is precisely that we should not draw any sharp distinctions between the natural environment made of living organisms (plants and animals) and the artificial environment we shape either by design or as a consequence of what we do with the designs we put out into the world. If we can appreciate the inherent value of that overall environment and the relations it depends on for its own sustenance, we can see that its constituent entities may be worthy of moral consideration independently of their usefulness to human welfare: We can thus include artificially intelligent entities in that group (comprising a growing range of entities), and to that end we need not necessarily rely on a standard list of properties such as sentience, consciousness, intelligence, or the ability to use a language.

[16]In addition, even if we stick to the notion of artificially intelligent beings as things, we could still ask deeper questions about them: In this way, as Holy-Luczaj (2015, 59–60) argues, we could "stop regarding them [things] as (easily) replaceable disposables," and "such a transformation [would] likely change the patterns of our consumption and thereby [have] a positive proenvironmental impact."

One reason suggesting that this may not be an appropriate set of metrics, or at least that the property-based approach may not work as a standalone solution, is that artificially intelligent entities may even *outstrip* biological entities in their capacities (such as the use of language), yielding the counterintuitive conclusion that they are worthy of even greater moral consideration than other beings in that respect. But the point here is not to rank different sorts of entities according to their degree of moral worth: It is rather to see whether they can be included as participants in our environment by looking at the role they play within that environment, and to see what moral consequences can be extracted on that basis.

That is why Deep Ecology seems to offer itself as an appropriate vantage point: It enables us to frame the moral and legal problem of artificial intelligence on an inter-relational approach closer to the kind of approach that has already been shown to work in tackling the great moral and political issues of inclusion and exclusion we have faced in the past. And it can do so drawing on philosophical insights from a broad range of inquiries. As Palmer (1998, 164) has noted, "a small and controversial new philosophical school [gains] revealing conceptual closeness to relatively illustrious philosophical ancestors," and that is precisely one of the aims of this paper: to show that environmental ethics in general, and Deep Ecology in particular, can draw on a broad range of insights from the past in dealing with a problem that is facing us now in the present and is poised to become even more pressing in the future. There are many aspects of environmental ethics and Deep Ecology that I do not discuss here, but I hope to have at least offered some good reasons for looking at artificially intelligent entities as entities forming part of a shared environment, for I submit that from this vantage point we can make some headway in dealing with some of the moral and legal issues their use and development might give rise to.

The idea of applying Deep Ecology to artificial intelligence runs parallel to the contemporary legal and ethical discussions on artificial intelligence and robotics: A debate is underway on whether to recognize electronic personhood for robots and whether the problem of their use can be managed within the current legal framework. It seems to me that before these questions can be given any definite solution, we need to challenge our assumptions. Naess (2008, 311), for example, argues that what we need is not a "shift from humans towards nonhumans, but an extension and deepening of care." This highlights a contrast between two different paradigms, suggesting that if we are to properly deal with artificial intelligence, we need to effect something along the lines of Naess's "substantial reorientation of our whole civilization" (ibid.).

Acknowledgements This paper is part of the project ALLIES (Artificially Intelligent Entities: Their Legal Status in the Future) that has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement n° 600371, el Ministerio de Economía, Industria y Competitividad (COFUND2014-51509) el Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander."

References

Armstrong, A.C. 2012. *Ethics and justice for the environment*. Abingdon: Routledge.

Baard, P. 2015. Managing climate change: A view from deep ecology. *Ethics & The Environment* 20 (1): 23–44.

Bedau, M.A., et al. 2000. Open problems in artificial life. *Artificial Life* 6: 363–376.

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

Coeckelbergh, M. 2009. Distributive justice and co-operation in a world of humans and non-humans: A contractarian argument for drawing non-humans into the sphere of justice. *Res Publica* 15 (1): 67–84.

———. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12 (3): 209–221.

Devall, B., and G. Sessions. 1985. *Deep ecology: Living as if nature mattered*. Salt Lake City: Peregrine Smith Books.

Farshchi, S. 2016. Let's bring Rosie home: 5 challenges we need to solve for home robots. *IEEE Spectrum* January 13. http://spectrum.ieee.org/automaton/robotics/home-robots/lets-bring-rosie-home-5-challenges-we-need-to-solve-for-home-robots/?utm_source=computerwise&utm_medium=email&utm_campaign=011916.

Floridi, L., and J.W. Sanders. 2001. Artificial evil and the foundation of computer ethics. *Ethics and Information Technology* 3: 55–66.

———. 2004. On the morality of artificial agents. *Minds and Machines* 14 (3): 349–379.

Fox, W. 2003. Deep ecology: A new philosophy of our time? In *Environmental ethics: An anthology*, ed. A. Light and H. Rolston III, 252–261. Oxford: Blackwell.

Gottlieb, R.S. 1999. Introduction. In *Ecology*, ed. C. Merchant, ix–xi. Amherst: Humanity Books.

Holy-Luczaj, M. 2015. Heidegger's support for deep ecology reexamined once again: Ontological egalitarianism, or farewell to the great chain of being. *Ethics & The Environment* 20 (1): 45–66.

Jonas, H. 1984. *The imperative of responsibility: In search of ethics for the technological age*. Chicago/London: University of Chicago Press.

Keller, D.R. 2008. Deep ecology. In *Encyclopedia of environmental ethics and philosophy*, ed. J. Baird Callicott and R. Frodeman, 206–211. Detroit/New York: Gale Cengage Learning.

———., ed. 2010. *Environmental ethics: The big questions*. Chichester: Willey-Blackwell.

Kellogg, P. 2014. Do machines have rights? ethics in the age of artificial intelligence. *Aurora*, http://aurora.icaap.org/index.php/aurora/article/view/92.

Keulartz, J. 1995. *The struggle for nature: A critique of radical ecology*. London/New York: Routledge.

Kortetmäki, T. 2016. Is broad the new deep in environmental ethics? A comparison of broad ecological justice and deep ecology. *Ethics & The Environment* 21 (1): 89–108.

Kurzweil, R. 1999. *The age of spiritual machines: When computers exceed human intelligence*. New York: Viking Press.

———. 2006. *The singularity is near: When humans transcend biology*. New York: Viking Press.

Naess, A. 1973. The shallow and the deep, long-range ecology movement: A summary. *Inquiry* 16 (1): 95–100.

———. 1986. The deep ecological movement: Some philosophical aspects. *Philosophical Inquiry* 8: 1–2.

———. [1989] 2001. Ecology, community and lifestyle: Outline of ecosophy. Cambridge: Cambridge University Press.

———. 2005. Theoretical dimension of deep ecology and ecosophy. In vol. 10 of *The selected works of Arne Naess*, ed. A. Naess, and A. Drengson, 546–550. Dordrecht: Springer.

———. 2008. *The ecology of wisdom: Writings by Arne Naess*. Berkeley: Counterpoint.

Nash, R.F. 1989. *The rights of nature: A history of environmental ethics*. Madison: The University of Wisconsin Press.

Palmer, C. 1998. *Environmental ethics and process thinking*. Oxford: Clarendon Press.

Rothblatt, M. 2014. *Virtually human: The promise—and the peril—of digital immortality*. New York: St. Martin's Press.

Tavani, H.T. 2011. *Ethics and technology: Controversies, questions and strategies for ethical computing*. 3rd ed. Hoboken: Willey.

Troster, L. 2008. Caretaker or citizen: Hans Jonas, Aldo Leopold, and the development of Jewish environmental ethics. In *The legacy of Hans Jonas: Judaism and the phenomenon of life*, ed. H. Tirosh-Samuelson and C. Wiese, 373–396. Leiden/Boston: Brill.

Yampolskiy, R.V. 2012. Artificial intelligence safety engineering: Why machine ethics is a wrong approach? In *Philosophy and theory of artificial intelligence*, ed. V. Müller, 389–396. Berlin: Springer.

Chapter 19
The Ethics of Choice in Single-Player Video Games

Erica L. Neely

Abstract Video games are a specific kind of virtual world which many engage with on a daily basis; as such, we cannot ignore the values they embody. In this paper I argue that it is possible to cause moral harm or benefit within a video game, specifically by drawing attention to the nature of the choices both players and designers make. I discuss ways in which games attempt to represent morality, arguing that while flawed, even games with seemingly superficial devices such as morality meters can attempt to promote moral reflection. Ultimately, I argue that the moral status of the actions depends on the effects of those actions on the player herself; if those actions make us less ethical then the actions are wrong. Unfortunately, it is not clear to me that players are always in a position to tell whether this is the case.

Keywords Immersive games · Virtual worlds · Moral choice · Intravirtual morality · Utilitarianism · Extravirtual harms · video games

19.1 Introduction

Video games have become ubiquitous in today's society, ranging from simple apps on a smart phone to immersive computer or console games that require 80 or 100 hours to complete. Furthermore, they are no longer the purview of a small fraction of the populace; they are a form of media that children grow up with and adults continue to engage with throughout their lives. As such, it is natural to wonder about the impact of games upon us: what kinds of effects are they having?

While much discussion in the popular press has been concerned with the effects of video games on children, we should not ignore their effects on adult players. While adults may be more morally developed and less easily influenced by the

E. L. Neely (✉)
Department of Religion & Philosophy, Ohio Northern University, Ada, OH, USA
e-mail: e-neely@onu.edu

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_19

messages in media, they are certainly not immune from them. I will argue that prominent accounts of ethics in video games, such as Miguel Sicart's (2009, 2013) ignore this fact by focusing too much on ideal players and not enough on actual players.

The increased focus on video games over the last decade dovetails with the attention many academics are devoting to extending principles of moral harm or benefit to virtual worlds. With the advent of online environments such as *Second Life* (Linden Research Inc. 2003), serious moral questions have been raised concerning the status of our actions in those realms. Can one cause harm via an avatar? Do our actions in a virtual world have moral status? There have been a variety of answers, but they all display a concern for the notion of causing harm within virtual worlds. Video games are a specific kind of virtual world which many engage with on a daily basis; as such, we cannot ignore the values they embody.

I will argue that it is possible to cause moral harm or benefit within a video game, specifically by drawing attention to the nature of the choices both players and designers make. For the purposes of this paper, I will set aside multiplayer games and concentrate on single player games. In such cases, we can separate the ethical consequences within the game from the consequences to the player. We can thus consider the ethical ramifications of actions from inside the game world and the relationship a player has to those actions; we can also consider the effects of the actions on the player herself. I discuss ways in which games attempt to represent morality, arguing that while flawed, even games with seemingly superficial devices such as morality meters can attempt to promote moral reflection. Contrary to Sicart, however, I believe that players are not always reflective about the moral choices they face. Ultimately, I argue that the moral status of the actions depends on the effects of those actions on the player herself; if those actions make us less ethical then the actions are wrong. Unfortunately, it is not clear to me that players are always in a position to tell whether this is the case.

19.2 Morality and Choices

Before diving into the details of how video games handle choices, one might wonder whether ethics is even relevant to this topic. I have argued elsewhere that moral standing is tied to having interests. (Neely 2013) These can range from very simple interests such as being free of physical pain to more complex interests such as those involved in our legal understanding of property ownership, however, if a thing, such as a rock, lacks interests, it is difficult to understand how one could either harm or benefit it. Within the realm of a single-player video game, one interacts with virtual characters; there are no other players, but there are other characters programmed into the game world. In one sense, those characters do not have interests, since they are not real – they are much like characters in dreams or fantasies. As such, it would appear at first glance that one could treat them however one wished: lacking interests, they also lack the ability to be harmed or benefitted, thus they seem to stand

outside of morality; one's actions towards them are neither morally praiseworthy nor morally blameworthy. Thus it may seem that there is not much to be said on this topic.

This is slightly hasty, however. Following Johnny Søraker (2012) we can distinguish intravirtual (inside the game world) and extravirtual (outside the game world) consequences of actions. From an extravirtual standpoint, video game characters, indeed, are fictional and thus cannot be harmed or benefitted extravirtually; any argument about morality must take another approach.[1] While we will consider this broader picture in a moment, let us first examine the former standpoint, i.e., the characters within their own context, as members of a particular virtual world.[2]

The ability to choose different actions has become an important part of many modern video games, and players expect the game world to reflect those actions. Games such as *Arcanum: Of Steamworks and Magick Obscura* (Troika Games 2001), *Dragon Age: Origins* (BioWare 2009), and *Mass Effect* (BioWare 2007) have offered players a multitude of possible actions, with different in-game consequences for each choice. In these games, actions towards the denizens of the game may have moral import because one's decisions have impact *within the game*. If the characters seem to be harmed (or benefitted) within the game world by your actions, then it is easier to attach moral standing to those actions. For instance, in *Arcanum*, the main character can choose to blow up a bridge leading to a particular town. At the end of the game, you discover that doing so causes the town to wither from lack of trade. It would appear, therefore, that your character has taken a morally wrong action – or at least one which has negative moral ramifications. On the other hand, if your character aided a person without any thought of gain, then you have likely done something virtuous.

In order to track the intravirtual moral consequences of our actions, many games have introduced systems that track the players' choices. I will now consider some of the ways in which intravirtual morality is handled, beginning with a fairly crude explicit system before turning to more complex instantiations of the system. While all of these systems have limitations, I will argue that they all permit an important type of moral exploration on the part of the player; there is thus a connection between the intravirtual moral consequences of the character's actions and the extravirtual moral exploration of the player.

[1] Of course, as Søraker notes, video games are particular states instantiated on physical devices and thus have an extravirtual component simply in terms of the bits on the machine; all of the characters, objects, and actions within the game thus have an extravirtual component in this sense. This is rarely the sort of extravirtual consequence we are concerned with from an ethical perspective, however.

[2] This is, presumably, the same sort of distinction we make for other art forms such as novels or films; on the one hand, it is false to say that Sherlock Holmes and Moriarty are enemies, since neither exist. However, in general when someone is making such a statement, they are actually talking about what is true within the fiction and, in this context, Sherlock Holmes and Moriarty are enemies. This distinction is discussed at length by Kendall Walton (1990) and is applied specifically to videogames by Grant Tavinor (2009).

19.3 Choices and Morality Meters

The idea that actions can have moral import within a game context is presumably the genesis of morality meters in video games. This is a fairly crude system for measuring morality. While there are variations, in general one extreme represents pure evil and the other pure virtue; the main character's morality is measured using this meter. Various actions will cause the meter to move incrementally in one direction or the other, depending on the scope of the action. A minor misdeed will make you only slightly less virtuous, while major scheming may cause the meter to drop significantly. We may call this a single-stream morality meter.

A serious issue with single-stream meters is that they display a single score to represent the player's morality – each action either is deemed morally good (adding points to the score), morally neutral (leaving the score unaffected), or morally wrong (subtracting points from the score.) This implies that enough morally good actions can cancel out a morally wrong action. Hence a player who performed an extremely evil action and then many extremely good actions to counter it would be viewed as no different than a character who has performed no evil actions and only a few small good actions. Yet one might well argue that the latter should be deemed morally superior to the former; at the very least, it seems there is a relevant difference between the two which is not captured by the game mechanics.

To address this concern, some games have separate scores to measure morally good and morally bad actions; we may call this a dual-stream morality meter. *Mass Effect* (BioWare 2007) and its sequels divided actions into two categories; a character could amass paragon points (if she performed a compassionate or heroic action) or renegade points (if she performed an apathetic or ruthless action.) For instance, when faced with the last surviving member of an alien species, choosing to set it free will earn paragon points while choosing to kill it will earn renegade points. In this way the designers ensured that one's actions never truly disappear; a character's new virtuous actions may outweigh his previous unethical actions, but they do not negate those actions. This is surely a more accurate representation of real world morality, since one's previous actions do not cease to exist simply because one has atoned: you may no longer steal, you may have repaid the person you stole from, but the fact remains that you once stole, and that cannot be undone.

There are large assumptions bound up in these meters, even if viewed only as intravirtual measures of morality. One critical problem is that they rarely take intent or context into account – all instances of X will drop or raise your morality by Y. Hence an accidental act is not distinguished from an intentional act, nor is there room for nuance; a poor character stealing bread because they are starving to death would be no different than a rich one stealing out of avarice.[3]

Another issue is that one may question the moral system underlying the meter. For instance, *Arcanum* contains a quest in which a farmer asks the player to kill

[3] As Heron and Belford (2014) note, this flaw generally rules out using Kantian ethics to measure morality in the game world, as there is no seamless way to determine the intent behind the actions.

some wild animals that are damaging his crops. If the player does so, her character's morality decreases and any good-aligned characters in her party will object. This supposes that killing these animals is an immoral act, which betrays an unfamiliarity or lack of care displayed for the amount of damage that vermin can do to crops. If the designers presented killing the animals as simply one of several ways of completing the quest, then perhaps this would be a plausible representation of morality; it could be the least virtuous way to achieve the goal. Since they did not, however, the moral message appears to be that allowing wild animals to ruin crops (and this farmer's livelihood) is more virtuous than removing those animals; this seems a rather questionable moral conclusion.

Morality meters, therefore, represent a particular view of morality within the game, and one with which the players may disagree. This is not in itself necessarily problematic. Grant Tavinor (2009) discusses the fact that players of a game are engaged in a kind of "make believe," in which we do not so much suspend our disbelief as agree to a set of fictions for the purposes of play. Thus when we play a game, one thing we do is engage with the game's world, which can include a particular moral stance.[4] Yet players will not always simply accept this stance uncritically, particularly if it does not seem well-supported by the rest of the game's fiction. In *Arcanum*, there is nothing to indicate that killing the animals should be seen as immoral, nor are there any other relevant experiences that would reinforce this message; this is a single instance of the moral situation, and it thus seems poorly motivated.[5] The morality meter seems, if not incorrect, at least debatable in its judgment of this instance.

Moreover, there is a very utilitarian feel about this assessment of morality. Single-stream morality meters, which simply adjust one way or the other due to your good and bad actions, represent an extremely simple hedonic calculus: if the amount of utility (positive morality points) outweighs the amount of disutility (negative morality points) then a character is good.[6] While dual-stream morality meters are somewhat more complex, they still seem largely consequentialist in character; awarding points based on each specific action, for instance, would not sit well with a virtue ethicist's idea that character is displayed through habituation, not single acts. A virtue ethicist approach simply does not fit well with an explicit morality meter, even though such meters are often presented as attempting to represent the character's moral character.[7]

[4]Sicart (2013) refers to this as being morally complicit with the game and its world.

[5]Sicart (2009) also discusses conflict between the rules of the game and the fictions of the game world, particularly when he discusses how the game *XIII* (Ubisoft 2003) portrays the character as a ruthless killer but the game will not allow her to kill police officers or innocents.

[6]Indeed, the entire scheme of awarding points is reminiscent of Jeremy Bentham (1823/1996), since actions which are more harmful or greater in scope do seem to award more negative points than those which have smaller consequences. It is not a perfect representation of his hedonic calculus, but it is in the same vein.

[7]Of course, this is not a truly utilitarian account of morality either, since it is relativized to the game world; in some sense, neither utility nor disutility is generated by an action, since the

A more fundamental objection to the idea behind morality meters is presented by Sicart (2009) when he argues that morality meters may have little to do with the player's ethical engagement, since they become just another mechanic to strategize over and manipulate. If a player knows that the game world will respond to him in certain ways if he takes certain actions, or if he crosses a certain threshold on the meter, then he may pay attention to the morality of his actions not for its own sake, but because he desires certain results in the game. This issue arises on multiple plays of a game, since one has an idea of what results will occur for certain actions based on past experience. However, many games have the ability to restore to a previous point via saving and reloading; this would enable a player to take an action, see what the effect is on her score, and redo it if she did not like that result.[8] While Sicart argues that such actions are purely strategic and devoid of moral reflection on the part of the player, I disagree. This, too, displays a kind of consequentialism: a player has her character take an action, evaluates the consequences, and then decides whether those are good consequences for the game *as the player wishes it to progress*. Admittedly, this represents a form of meta-gaming: the player is not necessarily concerned with the moral consequences as evaluated by the game. However, it enables the player to develop particular kinds of characters easily and see what happens to them within the game universe. This will not necessarily result in moral reflection on the part of the player, but it does not seem to prevent it either; the reflection simply will be over the character's actions/game as a whole, rather than over the consequences of a single action.

One way that games attempt to prevent this kind of meta-gaming is to attempt implementing more complex systems of morality. For instance, many games lack explicit morality meters but will alter the game world and people's reactions to you in response to what you do. This can be relatively simplistic; for instance, in *Arcanum* (Troika Games 2001), if a character is seen stealing, the town's guards will attack him. Alternately, the game can involve complex adaptations which are sensitive to dialogue and plot choices; in *Dragon Age: Origin* (BioWare 2009) there are many conversational paths with party members, and the dialogue choices a player makes will affect their attitudes toward her character. This is an attempt to display game-world consequences of one's actions in a less arbitrary fashion than through an explicit meter.

Such attempts can still be subject to Sicart's objection if they are too simplistic. For instance, if a particular dialogue seems to go poorly, a player may restore and try again. While I do not find his objection totally persuasive, as argued above,

actions are fictional. However, since such meters generally reflect what are considered good or bad consequences *within the game*, they are roughly utilitarian if one is engaged in the make-believe fiction of the world.

[8] Assuming that there is much of an effect on the gameworld; Heron and Belford (2014) criticize many implementations of morality meters because they are fairly shallow – the choices have few real consequences. This is an objection to how a system of morality is implemented in practice, however, rather than a fundamental objection to the idea of morality meters which Sicart appears to have.

his concern is further mitigated in some games by making the long-term effects of choices unclear.[9] One of the most interesting recent examples is in the game *Life is Strange* (Dontnod Entertainment 2015), which has a mechanic wherein the lead character can rewind time for short bursts, allowing her to try different options and see the results.

Three things make this mechanic particularly fascinating. First, the character is intensely self-reflective; in many situations, no matter what choice a player picks, the character wonders aloud whether she should choose the other. Unlike games with clear black and white paths, this leaves the player doubting and reflecting on his actions as well. Second, the rewind mechanic only works for a short period of time and does not continue indefinitely; once you have left one area and entered another, you cannot rewind past that point. Thus at some point one's choices are static – the player ultimately will have to make a decision and stick to it, unless she wishes to replay a large portion of the game.[10] Since many of the choices have long-term consequences, the player can pick what *seems* best, but he may be wrong about whether that choice actually *is* best. Third, partway through the game the character starts losing the ability to rewind time in some situations. This lends an unexpected urgency to dialogue and action choices in those cases – when the character is faced with trying to talk someone out of committing suicide, knowing that you cannot rewind makes the player's choices feel more significant. The fact that the game explicitly built in the players' ability to try different options and then took it away lends a weightiness to the consequences beyond what typically seems to be present in video games.[11] These factors combine to make the game world's adaptation to a player's choices extremely compelling and promotes a greater thoughtfulness with regard to moral decisions than most games.[12]

One of the interesting aspects of *Life is Strange* is how wildly unrealistic its implementation of moral choice is; in real life we cannot try out different options and rewind to see what would happen if we tried another path. In general, while morality meters are fairly crude devices, they are attempting a fairly realistic representation of morality: just as we judge people by their actions in the real world, the designers attempt to do so in the game world as well. These systems have limitations – most of us view morality as slightly more complex than simply

[9] Sicart (2013) looks at this in greater detail, particularly praising *Fallout 3* as an example of a game which does this well.

[10] Unlike many games which allow a player to save whenever he wishes, *Life is Strange* only allows saves at particular checkpoints; to change options after the rewind window closes, a player would have to reload to the previous checkpoint and play the game through to that dialogue or action choice again.

[11] Once again, this is reinforced by the fact that saving and reloading the game is somewhat constrained and thus adds a price to deciding to change one's choices.

[12] This is in part because *Life is Strange* has a stronger narrative than many games due to its linear nature and way of handling player choices. While I agree with Tavinor (2009) that frequently games have difficulty with narrative due to gameplay constraints, *Life is Strange* uses moral choices to reinforce different narrative possibilities in an extremely effective manner.

reducing a person to a number or pair of numbers, and we cannot generally engage in the sort of meta-gaming that the ability to save and reload allows. Yet despite these limitations they still can promote moral thinking. Moreover, *Life is Strange,* which explicitly embraces some of the artificialities of typical play by incorporating it into the story line, demonstrates that even a wildly artificial system does not preclude such deliberation.

Having said that, the way in which the moral thinking occurs will likely differ depending on how obvious or artificial the system is. Attempts to modify the player's experiences based on his actions in the game clearly is a reflection of what happens in the real world. Our actions have consequences; the world (and people in it) respond to what we do. There is a need for some system of in-game morality if the game world wishes to seem realistic; in general, a world in which observed stealing has no consequences is not convincing.[13] Similarly, it is easier to be immersed by a world where not all actions are presented as having the same moral ramifications. The morality meter or adaptation reinforces the fiction of the world.[14]

The attempt to make a convincing game world has interesting consequences, as our identification with our characters affects what we are willing to do with them. Michael Nagenborg and Christian Hoffstadt (2009) noted that the more a player sees her avatar as a reflection of herself, the more her own ethical code comes into play.[15] If she strongly identifies with a particular character in a game, she will be less willing to have that character commit actions she views as morally wrong; if she does not strongly identify with that character, then she is more likely to pay attention to the fictional nature of the game and thus feel that any action is morally acceptable (since, after all, the action is not truly occurring.)[16] A sufficiently immersive game world, then, has the potential for prompting moral deliberation. A player may not see his avatar as a perfect reflection of himself, retaining his own moral code. However, if he sees his character as embodying particular traits, then he may react as he believes such a person would react. In this case, he is not seeing all actions as permissible; he is instead approaching the scenario from a particular moral standpoint, albeit not the same one as he likely has in the real world.[17]

[13]Presumably even if a game is set in a lawless dystopia, people will be annoyed if you take their belongings.

[14]Note that by "immersed" I simply mean that a player is deeply mentally engaged with the game, much in the same way that one can be drawn in by the fiction of a book or movie. Many games attempt to create worlds that promote this by trying to be relatively realistic (insofar as their setting allows).

[15]Although I would note that some research (Lange 2014) suggests that the majority of players engage with moral choice systems using their own moral code regardless of how much they identify with a character.

[16]Note, with Gorrindo and Groves (2010), that what we do with our avatars is not literally what we are willing to do in real life; the fact that you are willing to murder someone in a game does not imply you would murder in real life. Your avatar's actions may are not a literal map of your actions – they at best provide insight into your personality.

[17]It will be interesting to see how this evolves as we have more immersive virtual worlds – will players be less willing to choose the "evil" path in a game? Will there be a point at which it simply

It is not clear to me that this kind of immersion is always required, however. As my response to Sicart on morality meters indicates, I believe that players will sometimes engage in meta-gaming to aim for a particular kind of game experience. Similarly, games such as *Life is Strange* use the artificial nature of the game to allow for a greater freedom to explore options than real life allows. I do not necessarily regard this as ethically inferior to a game in which a player is more directly immersed (or where the moral system is less obvious). Rather, I believe they promote different kinds of potential ethical experiences. A game in which a player strongly identifies with a character will engage her ethically at each decision point; she may agonize over what to do in various situations because her avatar is an extension of herself and thus the choice seems more real. However, when a player is engaged in meta-gaming, there is still the potential for moral evaluation. That evaluation, however, is more likely to be of the ultimate experience of the game as a whole: if I pick choices X, Y, and Z, did the game react in a convincing or satisfying way? The player's character is thus much closer to a character in a book or a movie, but one which the player directs – the player makes choices, but there is little identification with those choices. It is thus about the particular experience of the game as a whole.

19.4 Extravirtual Harm

This distinction between evaluating one's actions in the game and evaluating the game experience as a whole brings up larger questions of morality. It seems clear that, within a game world, one can take ethical or unethical actions; there are ways of harming or benefitting characters inside the game context. However, this leaves open the larger question of whether you are causing moral harm or benefit outside of the game world; are there extravirtual consequences of your actions?

This issue is frequently framed in terms of whether it is morally problematic to play violent video games. As Matt McCormick (2001) notes, it has become common for the media to connect video game playing to events such as mass murders and school shootings; it is almost stereotypical at this point to reveal that such perpetrators loved playing first-person shooting games. Even without that connection, some games are extremely brutal or gruesome, and many wonder whether there is something unethical about engaging with them. We can thus raise questions on both a micro and macro level: is it wrong to commit actions in a game if we would deem those actions wrong in the real world? Is it wrong to play a game which encourages such actions? Or, should we argue with Sicart (2009, 2013) that players are sufficiently capable of moral reflection and thus are not susceptible to

becomes too realistic to maintain a separation between their own morality and the game's morality? Or will we become gaming chameleons, wherein we can successfully inhabit a range of moralities, depending on the character we are playing?

being morally harmed by games? The truth, I will argue, lies somewhere between media hysteria and Sicart's blithe assurances of moral reflection – while gamers are capable of moral deliberation, it is not clear to me that they always engage in it.

Let us consider a somewhat fanciful example. *World of Warcraft* (Blizzard Entertainment 2004) contains a quest in which you are instructed to take a sharp stick and poke baby monkeys to cause them distress.[18] Within the game context this action is essentially seen as a necessary evil – the fact that you are asked to do this by a particular faction is motivation to later repudiate that faction. However, since generally we frown on torturing animals in the real world, one might wonder whether this quest is wrong to undertake in some larger sense.

In order for our actions to cause moral harm, someone's interests must be harmed. From an extravirtual perspective, clearly we cannot claim that the monkeys are actually harmed since they do not exist. The only existing entity directly involved in the scenario is the player; as such, it appears that the only being who could be harmed is that player. The question then becomes whether a player is somehow causing harm to herself by engaging in the action. This is a virtue ethics approach which addresses the effect on a player's moral character; if by performing the game action, the player is apt to become less ethical in real life, then the action is wrong to take within the game.[19] In essence, the player is rendering herself less virtuous by taking that action, and thus indirectly could be promoting future harms to others. For instance, if repeatedly engaging in violent activities in a game is rendering the player less sensitive to the effects of violence on others, then she should refrain from those activities. Members of society have an interest in adhering to the ethical standards of that society; a choice which makes one less likely to have empathy for others in the social group is impeding one's social interests.

However, it is not clear whether these actions will translate into future harms. McCormick (2001) and Coeckelbergh (2007) each reject utilitarianism and deontological ethics in this regard because there is not enough evidence to connect video game playing to bad future actions.[20] Yet it is not clear that they establish that harm to one's character actually occurs in playing these games. Coeckelbergh claims that

> The more precise conditions for a game to be morally problematic are not only (1) that there is violent content, but also (2) that there are particular structural similarities between the virtual and the real world in place, and (3) that they un-train – or, at least, do not allow or inhibit development and training of – empathy.[21] (p. 227)

[18] While *World of Warcraft* is a multi-player game, this particular example does not involve any multiplayer elements and thus is akin to a quest in a single player game.

[19] I am far from the only person to suggest this approach. For instance, McCormick (2001) raises this as a possibility and Mark Coeckelbergh (2007) develops it further.

[20] Indeed, the empirical studies are decidedly mixed in their results, and I tend to agree with Coeckelbergh's assertion that "philosophers are tempted to pick out the one or few [empirical studies] that suit their arguments best." (Coeckelbergh, 2007, p. 220)

[21] This could be somewhat too restrictive if, in fact, there are non-violent actions which also negatively affect moral character. Such actions were beyond the scope of Coeckelbergh's argument, but a broader use of his definition may require an expansion of this clause.

Clearly, not all seemingly unethical game actions will translate into real world harms. Some actions may be neutral in their effect on the player. For instance, consider a player who steals in a video game. Since frequently games will allow rogues or thieves as characters, this player may see these behaviors as tacitly endorsed within the game. As such, he may see his behavior as divorced from the real world: he can simultaneously see stealing as wrong in this world while believing it permissible in the game world. Assuming he is able to distinguish the two worlds, these actions are not apt to make him less ethical. In this case, even if there are some structural similarities between the virtual and real world, they are sufficiently different to allow for ethical distinctions.

Similarly, some games deliberately encourage thinking about ethical dilemmas and wrestling with what actions to take. As mentioned before, *Life is Strange* (Dontnod Entertainment 2015), delays the appearance of many consequences in a way that lends significance to player choices and encourages players to try different paths and see what happens. When thoughtfully done, this kind of experimentation can be morally beneficial to the player – not only may it fail to make her more unethical, it may instead aid her moral development by increasing her sensitivity to ethical choices and their ramifications. So actions, even unethical actions, could increase empathy.[22]

Thus, with respect to the aforementioned *World of Warcraft* quest, poking baby monkeys with a stick in the game is not necessarily wrong, assuming the player is not thereby more likely to commit harm in real life. If, say, she takes the quest and experiences moral revulsion while performing it, the quest may instead be morally beneficial; she has learned something about her reactions to torture or animal cruelty. This is a kind of philosophical thought-experiment in video game form; while the trappings may be fantastic, the moral dilemmas faced in games can reflect larger ethical questions about the treatment of animals, the lengths one should go to in order to appease authority figures (such as quest givers) and so forth.

Much of Coeckelbergh's attempt to lay out conditions seems quite convincing, therefore. Yet the problem remains that it is fairly abstract – he has argued that actions are wrong to take if they make one less ethical by inhibiting empathy, but he has not said how to determine whether this is so. Perhaps in the monkey example it seems likely that the player is not harming her moral character, since she experiences an appropriate reaction. But what about an instance to the contrary? What if the player believes himself capable of divorcing the video game from reality but, in fact, is being influenced by it and is acting less ethical in the real world? How can one tell that the action is wrong to take?

There are two things to note in response to this objection. First, this raises interesting questions about distinguishing the game from reality, and I believe that the game context itself is relevant to this; I thus believe that Coeckelbergh's second criterion is useful here. Games which mimic reality are relevantly different from

[22]This is presumably part of what Sicart (2009, 2013) finds promising about the creation of ethical video games.

games which take place in vastly different worlds. If a game is set within a fantasy world wherein a player is a wizard casting spells and slaying dragons, there is probably a sufficient disconnect between that world and this one to render it easy to distinguish the two; no matter how tempting it may be to fling a fireball into the middle of a boring faculty meeting, one is aware that this is not possible. On the other hand, a game such as *Grand Theft Auto IV* (Rockstar North 2008) involves situations which occur in the real-life. The chances are thus higher that such games will cause moral repercussions for the player due to the direct parallels between the actions in that game and actions in the real world.[23]

Second, there is an important distinction between the wrongness of an action and our being able to determine that wrongness. The former, more theoretical question, is the one which Coeckelbergh and I have been addressing; the latter is the pragmatic question of how to act upon that theoretical result. While fairly convinced by Coeckelbergh's proposed answer to the former question, I find the latter more troubling. It is true we can study general effects of video games upon individuals to see whether there are trends in what kinds of games and actions have good or bad effects upon the players and their future actions.[24] However, there are currently contradictory studies (as noted above), and I do not know whether this situation will improve. If it does not, then we have little way of telling what the effects on a player's character are.

This is particularly troubling given the tendency among some writers, Sicart (2009, 2013) in particular, to overstate the moral reflection among gamers. This is likely in response to the popular portrayal of gamers as being almost passive puppets in the hands of violent video games, shaped into hateful, violent beings through playing first-person shooting games. That is clearly a caricature of gamers and their responses to games. Yet, Sicart risks swinging too far the other way when he notes that "When I write about players, I am referring to an implied, model player . . . who has experience playing games and has the ethical maturity to understand them as an expressive medium." (Sicart 2013, p. 25).

I agree that, in general, a "player is a moral user capable of reflecting ethically about her presence in the game, and aware of how that experience configures her values."[25] (Sicart 2009, p. 17) However, I am not certain that adults are as immune to influence as Sicart believes. While adults do usually have more experience with moral thinking than children, I do not agree that our morality is fully-formed and unchanging; indeed, if games can promote virtue and moral thinking, as many argue, then they can do the reverse as well. One cannot be susceptible to virtue unless one is also susceptible to vice. Furthermore, while players are clearly *capable* of moral

[23]Note that this also increases the possibility of moral benefit, not simply moral harm.

[24]This is already being done by researchers such as Saleem, Anderson, and Gentile (2012).

[25]Note that we are setting aside the question of child players here – both Sicart (2009, 2013) and Tavinor (2009) explicitly distinguish players who are not adults and thus not morally-formed in order to argue for age-restrictions on games. Adults seem to be viewed as having a stronger moral center and as being more capable of reflection.

reflection, this does not imply that they always *engage* in moral reflection. As such, there continues to be a risk to actual players, even if there is no risk to the theoretical player.

In particular, while a single action seems unlikely to change the moral character of a gamer, it is less clear that a pattern of actions will have no effect. It may well be that completing one morally dubious action in a video game or even playing one morally dubious video game will not significantly affect one's character. This does not imply, however, that repeating the actions has no effect. Exposure to one idealized body image is unlikely to cause an eating disorder, yet cumulative exposure has a much greater chance. (Stice et al. 1994) A similar effect may be true for video games. Perhaps it may not matter if one takes violent actions in a single game, but it may matter if it is part of a greater trend. Similarly, playing a single first-person shooting game where the hero is white and the targets are all non-white may not affect one's character, but perhaps playing many such games does.[26]

Thus, despite being sympathetic to Sicart's emphasis on the reflective potential of gamers, I am less optimistic about its practical value. The fact that we are capable of reflecting on our choices does not prevent us from making harmful choices, even if we are not aware of it. Thus choices within a game may lead to extravirtual harm, not necessarily in an overt fashion, but by subtly influencing us.

19.5 Conclusion

With the increasingly pervasive reach of video games, it is important to consider their moral ramifications. I have argued for an emphasis on choice as a way of understanding various ethical issues that arise in this arena. In single-player games we must distinguish between the intravirtual effects on the game world and extravirtual effects on the player of the game. A desire to adapt the game world to player choices has, in part, led to the advent of morality meters and other systems of morality tracking within the game. Unfortunately, these have certain limitations. In particular, their inability to consider the nuances of a particular action is problematic, as is the question of what system of morality is in play. Breaking with Sicart (2009, 2013), however, I am less concerned with the idea that players could strategize to obtain particular results; this strikes me as simply another venue for potential ethical reflection. Thus I believe that even explicit or obvious systems of morality in video games may be useful for ethical reflection.

The idea that good actions could cancel out bad actions has caused many to seek alternate ways of portraying morality within video games, whether through separate meters tracking good and bad actions, or simply through adapting the game

[26] Ultimately I think these kinds of concerns bind game designers as much as players, since building a world that encourages certain kinds of actions may make players less inclined to deliberate on the worth of those actions and thus less inclined to see how their choices are affecting their values.

world without any explicit measure of morality. Once players have real choices within the game world that world must reflect those choices or else it lacks realism. Interestingly, that realism has consequences for a player's actions within the game; the more she identifies with a character in a game, the less she is willing to use that character to violate her own sense of morality. Immersion is not required for ethical reflection, however, as the act of experimentation within a game world can also lead to reflection on the part of the player. This is true even when the moral system is implemented in an extremely artificial way, such as in *Life is Strange*. (Dontnod Entertainment 2015).

Outside of the game world, we must consider the effect of video game actions on a player; in particular, we must ask whether the actions can cause moral harm to that player by rendering him less ethical. Unethical actions in a game do not necessarily have this result, as the player may be able to separate actions in the game from actions in the real world. Furthermore, many games with sophisticated conceptions of morality specifically encourage the player to deliberate among the possible choices; this deliberation may aid our moral development. Unfortunately, it is not clear whether we will always be able to tell if a game is harming us; while ideally players will engage in self-reflection, actual players do not always do so and may not be as capable of moral deliberation as the ideal player.

References

Bentham, J. 1996. In *An introduction to the principles of morals and legislation*, ed. J.H. Burns and H.L.A. Hart. New York: Oxford University Press.

BioWare. 2007. *Mass effect*. Microsoft Game Studios.

———. 2009. *Dragon age: Origins*. Redwood City: Electronic Arts.

Blizzard Entertainment. 2004. *World of Warcraft*. [Online Game] Blizzard Entertainment, played 6 July 2016.

Coeckelbergh, M. 2007. Violent computer games, empathy, and cosmopolitanism. *Ethics and Information Technology* 9: 219–231.

Dontnod Entertainment. 2015. *Life Is Strange.* Square Enix.

Gorrindo, T., and J.E. Groves. 2010. Crime and hate in virtual worlds: A new playground for the Id? *Harvard Review of Psychiatry* 18 (2): 113–118.

Heron, M.J. and P.H. Belford. 2014. Do you feel like a hero yet? Externalized morality in video games. *Journal of Games Criticism* 1(2). Retrieved from http://gamescriticism.org/articles/heronbelford-1-2

Lange, A. 2014. "You're just gonna be nice": How players engage with moral choice systems. *Journal of Games Criticism* 1(1). Retrieved from http://gamescriticism.org/articles/lange-1-1

Linden Research Inc. 2003. *Second life*. [Online Game] Linden Research Inc. Accessed 6 January 2016.

McCormick, M. 2001. Is It wrong to play violent video games? *Ethics and Information Technology* 3: 277–287.

Nagenborg, M., and C. Hoffstadt. 2009. A life no longer worth playing: Some remarks on in-game suicide. *Journal of Gaming & Virtual Worlds* 1 (2): 83–95.

Neely, E.L. 2013. Machines and the moral community. *Philosophy and Technology* 27 (1): 97–111. https://doi.org/10.1007/s13347-013-0114-y.

Rockstar North. 2008. *Grand Theft Auto IV*. Rockstar Games.

Saleem, M., C.A. Anderson, and D.A. Gentile. 2012. Effects of Prosocial, Neutral, and Violent Video Games on Children's Helpful and Hurtful Behaviors. *Aggressive Behavior* 38: 281–287.
Sicart, M. 2009. *The ethics of computer games*. Cambridge: MIT Press.
———. 2013. *Beyond choices: The design of ethical gameplay*. Cambridge, MA: MIT Press.
Søraker, J.H. 2012. Virtual worlds and their challenge to philosophy: Understanding the "intravirtual" and the "extravirtual". *Metaphilosophy* 43 (4): 499–512.
Stice, E., E. Schupak-Neuberg, H.E. Shaw, and R.I. Stein. 1994. Relation of Media Exposure to Eating Disorder Symptomatology: An Examination of Mediating Mechanisms. *Journal of Abnormal Psychology* 103 (4): 836–840.
Tavinor, G. 2009. *The art of videogames*. Malden: Wiley-Blackwell.
Troika Games. 2001. *Arcanum: Of Steamworks and Magick Obscura*. Sierra Entertainment.
Ubisoft. 2003. *XIII*. Ubisoft.
Walton, K. 1990. *Mimesis as make-believe*. Cambridge: Harvard University Press.

Part VI
Trust, Privacy, and Justice

Chapter 20
Obfuscation and Strict Online Anonymity

Tony Doyle

Abstract The collection, aggregation, analysis, and dissemination of personal information permit unnerving inferences about our characters, preferences, and future behavior that were inconceivable just a couple of decades ago. This paper looks primarily at online searching and the commercial harvesting of personal information there. I argue that our best hope for protecting privacy online is anonymity through obfuscation. Obfuscation attempts to throw data collectors off one's digital trail by making personal data less useful. However, anonymous web searching has costs. I examine two of the most serious and urge that they are worth paying in the light of the heavy toll the commercial gathering and analysis of our information takes on privacy and autonomy. I close with some thoughts on (1) how individual, rational decisions have led to a surveillance regime that few would have chosen beforehand and (2) the alleged autonomy of information technology.

Keywords Privacy · Big Data · Predictive analytics · Anonymity · Obfuscation · Autonomy · Welfare

20.1 Introduction

The analysis or mining of big data has delivered many benefits. It has been a boon for detecting credit card fraud and money laundering, monitoring traffic flows, aiding learning, refining translation, and tracking public health trends and threats, among other things (Aquisti 2014, pp. 76–77; Barocas & Nissenbaum in Lane 2014, p. 44; Mayer-Schönberger and Cukier 2013, pp. 34–35 and 115; Schneier 2015, pp. 8, 27–28, 136–37, and 235). Take public health. Finding that a significant number of people in a given area are suddenly seeking information on, say, flu symptoms can

T. Doyle (✉)
Hunter College Library and Hunter College Philosophy Department, New York, NY, USA
e-mail: tdoyle@hunter.cuny.edu

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_20

signal an impending epidemic and can give public health officials far timelier data than conventional methods could (Aquisti 2014; Mayer-Schönberger and Cukier 2013).

However, big data is bearing down on privacy. Plummeting storage costs mean that vastly more personal information is gathered now than in pre-digital days. Also, privacy was once protected by the scattering of personal information across a sprawling landscape of file cabinets, archives, and stand-alone computers (Rule 2007). Now that information is available from a single point. Once aggregated, personal information can be analyzed to yield rich portraits, which in turn permit uncanny inferences about our activities, preferences, and future behavior. This windfall can then travel the world, flouting the time-honored norms that had previously governed information flows (Nissenbaum 2010).

How should we respond? I argue for obfuscation-supported anonymity online. Obfuscation attempts to throw third parties off one's digital trail by "producing misleading, false, or ambiguous data with the intention of confusing an adversary or simply adding to the time or cost of separating good data from bad" (Brunton and Nissenbaum 2011). I focus on the commercial surveillance of our web searching. I accept Brunton and Nissenbaum's conclusion that the commonly offered solutions to mass commercial surveillance, informed consent and legislation, fail to protect our privacy adequately (Brunton and Nissenbaum 2011). I see obfuscation online as the digital equivalent to wearing disguises or speaking low so that our neighbors cannot hear. I call for a return to something approaching analog age levels of privacy. My argument goes like this. Obfuscation promotes anonymity; anonymity promotes privacy; privacy promotes autonomy; and autonomy promotes well being by enabling us to increase our opportunities and advance our projects (Tavani and Moor 2001). Commercial tracking, monitoring, and profiling are bad insofar as they tend to be inimical to privacy and thus to autonomy.

After acknowledging some costs of obfuscation, I close with a discussion of two proposed threats to privacy. The first is what Alfred Kahn calls the tyranny of small decisions (Kahn 1966). Consumers can sometimes make a series of decisions which, although rational from the perspective of self-interest, can add up to a result that almost no one would have chosen at the outset. Privacy is plausibly a victim the tyranny of such small decisions. Second, I discuss the proposal that information technology is an autonomous force that human agency is powerless to affect and that privacy is an inevitable casualty of rampaging technology. I reject this notion.

20.2 Big Data's Revelations: Some Examples

Big data's successes come from using data collected for one purpose and applying it to other, apparently unrelated, purposes. This process enables data collectors to discriminate ever more finely among people to arrive at the optimal decision, from the data holder's point of view, about how to treat a given person at a certain time (Rule 2007).

Since the early 1990s, insurers have used credit scores to determine who to write policies for and what to charge for the policies they do write, since people with bad credit are significantly more likely to make claims than those with good credit (Rule 2007). More recently data miners have honed their techniques, revealing, for instance, that folks who buy cheap motor oil, Chrome-Skull car accessories, hang out in the local dive, or have friends who meet these criteria, tend to have bad credit and presumably are bad insurance risks as well. By contrast, those who buy home carbon monoxide sensors, snow roof rakes, felt "feet" for their furniture, and premium bird seed almost always pay their bills on time (Duhigg 2009; see also Mayer-Schönberger and Cukier 2013).

The apparently innocuous data that we generate as we go through the motions is more or less up for grabs, and in critical mass it enables miners to categorize us according to race, ethnicity, and sexual preference, as well as according to more specific criteria like *gambler*, *smoker in the house*, *adult with elderly parents*, and *adult with wealthy parents* (Singer 2013). The categories people are placed in can affect the products and prices they are offered, the quality of service they receive in a call center, the ads they see online, or whether they are bypassed altogether. This is the panoptic sort that Oscar Gandy (1993) presciently warned about long ago. The techniques of big data permit the sorting of individuals based on many criteria, chief among them "their estimated presumed economic or political value" (Gandy 1993, p. 1). Big data's ability to do so has improved immensely over the years, thanks to dramatically reduced storage costs, greatly expanded networks, and ever more sophisticated techniques of analysis, from which novel, precise, and profitable inferences can be made about data subjects. This elaborate process enables data collectors to determine the goods and services that people will be offered in a way that serves the interests of the collectors and their clients. Gandy calls the panoptic sort a "difference machine," a "discriminatory technology," that "allocates options and opportunities" on the basis of personal characteristics and how people are likely to act (Gandy 1993, pp. 15 and 17). The sort is "an integrated system that is involved in the identification, classification, assessment, and distribution of individuals to their places in the array of life chances" (Gandy 1993, p. 35). Nearly all of this happens without people's awareness about what is collected, who it is being shared with, or what those third parties are doing with the information once they have it. (Gandy 1993, p. 54).

Again, big data is all about effective discrimination: Businesses want to know both who to seek out and who to avoid. The reward for effective discrimination among retailers is increased profit (Rule 2007; Schneier 2015). For instance, those who frequent gambling sites might be a bad risk for a bank loan (Steel and Angwin 2010). More subtly, a detailed picture of one's health can emerge without any third party access to one's medical records. Obesity, which stands in for a suite of health risks, can be reliably inferred from the following: regular fast food dining, frequent online shopping for clothes, being a childless minivan owner, and subscribing to premium cable (Walker 2013). One data broker was able identify people who were probably arthritic by looking at cat ownership, preference for jazz, and participation in a sweepstakes (Walker 2013). Risk for high blood pressure, diabetes,

and depression can be inferred from people's hobbies, the websites they visit, and how much TV they watch (Mayer-Schönberger and Cukier 2013). The same goes for race. Zip code or mother's level of education are effectively stand-ins for race (Ohm 2014). The canny third party need not have any information specifically about our own characteristics. Information about others who are relevantly like us suffices to sort us in all kinds of ways, even if we have not consented to the release of the critical information in question (Barocas and Nissenbaum 2014).

20.3 Some Concepts

Why value privacy? It tends to promote autonomy (see Cohen 2000). Autonomy means being able to choose, free of coercion and manipulation, in the light of one's own considered conception of the good life. Maximum autonomy is nice, but I deny that it is intrinsically valuable. Take two worlds that are equal in well-being or the satisfaction of people's preferences or desires, say ours and Huxley's brave new one. I will assume that there is no moral basis for preferring one to the other. Autonomy and the privacy that tends to promote it are nothing but interests or informed preferences. They are not worth safeguarding for their own sakes. It follows from this that I deny any right to autonomy or to privacy (compare Nissenbaum 2010). Nevertheless, people are generally better off when they have more rather than less autonomy. Same for privacy. By threatening privacy, information technology limits autonomy and undermines welfare. Privacy counts a lot.

Enter anonymity and obfuscation. Namelessness does not suffice for anonymity (Wallace 1999). Thanks to sophisticated data analysis techniques, anonymity now requires more than blocking third-party access to unique identifiers like a social security number or information about one's movements over time. Following Kathleen Wallace, I will assume that social inaccessibility is needed for anonymity (Wallace 1999). Social inaccessibility obtains when "others are unable to relate a given feature of the person to other features" (Wallace 1999, p. 24). For instance, when I walk into a Starbucks in a strange town and pay with cash, I assume that no one is able to relate my appearance, the sound of my voice, or how much change I slip into the tip jar, with other information about me like my name, my credit rating, or my fondness for corn flakes. For others in the shop I am "unreachable" or "out of grasp" (Nissenbaum 1999). Or take Satoshi Nakamoto, the reputed creator of bitcoin (Throsby 2015). Assume that *Satoshi Nakamoto* is a pseudonym and that not even his partner, family or closest friends can connect him with bitcoin. He has isolated this trait from his other traits (Wallace 1999). By thus shrouding his identity he is anonymous as the creator of the virtual currency.

Privacy and anonymity are different. A person can enjoy privacy while lacking anonymity and retain anonymity while losing privacy. Suppose that the president of the United States has absolutely no anonymity: Everyone knows who he is in all his guises. Wherever he goes, whenever he speaks, he is instantly recognized by all who see or hear him. Yet he still has privacy when he retires to the White House

residence at the end of his long day. By contrast, my anonymity, but not my privacy, can remain intact even if a voyeur lurks as I try on new clothes in a changing room at the local department store. Nevertheless, when it comes to data, robust anonymity goes a long way towards protecting privacy or blocking unwanted flows of personal information.

Enter obfuscation. It can help promote online anonymity and thus privacy and autonomy. The "strict" in my title signals that old-school anonymity is no longer up to the task of keeping us out of reach from profit-driven, data-hungry third parties. Big data's techniques can often "de-anonymize" allegedly anonymized data, "using information that nobody would classify as personally identifiable" to disclose identity (Ohm 2010, p. 1704). Back in 2000, Latanya Sweeney found that 87% of Americans could be identified by a combination of their five-digit zip code, gender, and date of birth (Schneier 2015, p. 44). Also, our searches leave "data fingerprints" from which our identity can be inferred or guessed at with reasonable certainty (Ohm 2010, p. 1723). A much discussed early case involves the release of apparently anonymized AOL searches covering 3 months. Intrepid reporters at *The New York Times* were able to identify some subscribers on the strength of so-called vanity searches and because searches were rich in revelatory locations (Barbaro and Zeller 2006). Examples are legion. (See for example Ohm 2010, 2014; Schneier 2015; Zimmer 2010.)

Obfuscation confounds data gatherers by making the data itself ambiguous, harder to use, and thus less valuable (Brunton and Nissenbaum 2011, 2013). The technique is probably as old as life itself. Consider the monarch and viceroy butterflies. As a result of feeding on milkweeds as larva, monarchs are toxic to many vertebrates (Oberhauser 2011). The species advertises its venom in striking black and orange. A bird that has tried to feed on a monarch in the past will presumably remember the unpleasantness and avoid similarly colored butterflies in the future. It is even possible that natural selection favors predators that resist preying on monarchs at all. At least one mimic has capitalized on the monarch's combination of vibrant coloring and revolting taste: the viceroy. Today the non-toxic viceroys are all but indistinguishable from their noxious cousins. It is easy to see why natural selection might promote obfuscation here. For the predator, information about the potential quarry is ambiguous. Is the brilliantly arrayed insect ahead a hearty lunch or a possible last meal? The shrewd bird will avoid anything like a monarch. The situation is worse if natural selection fashions still more mimic species. Online obfuscation works similarly, attempting for instance to disguise the surfer's identity or the nature of her queries enough to throw unwelcome third parties off the trail. The point is to drown the signal out with ever more noise (Howe and Nissenbaum 2009; Brunton and Nissenbaum 2011).

Anonymity through obfuscation is justified. Again, we do not know who is getting what about our internet activity, what they are doing with it, or what happens to us as a result (Nissenbaum 2010). Moreover, users are generally not in a position to negotiate whether their information will be taken up in the first place, let alone what happens thereafter (Aquisiti 2014). If I am justified in disguising my appearance in public, particularly in the light of ever more common

face-recognizing surveillance, then surely I am justified in obfuscating my online behavior to avoid monitoring and profiling. Until data collectors or regulators can guarantee strict anonymity or confidentiality, people have little choice but to obfuscate.

20.4 Costs of Obfuscation

The defender of obfuscation needs to acknowledge that the practice has costs. One cost is free ridership (Brunton and Nissenbaum 2011). We pay for ostensibly free internet services like search engines, informational websites, apps, and social media in the coin of personal information. In fact, when we do a search in Google, our information is the product. The real customers are advertisers (Schneier 2015). Since obfuscators are not paying this price, they saddle everyone else with the costs of monitoring, profiling, and targeting, while reaping the benefits of a genuinely free internet (Brunton and Nissenbaum 2011).

However, as I mentioned above, generally if we want a given product or service online, we have little choice but to part with personal information. Again, most of us will likely never be in a position to know just how we are affected by the information skimmed from our clickstreams. When shopping online, not only are our purchases duly noted, so too is our browsing. Most people would not, at least for the time being, tolerate anything like this degree of monitoring as they wandered through the aisles in a store (Angwin 2014; Schneier 2015). Of course, data holders can always object that, if people don't like it, they can just head on down to their neighborhood big box or mom and pop.

Maybe, but the future does not bode well for shopping offline. Take music and books. Determined shoppers can probably still find most of what they want in a book or record store, at least in big cities. However, these days are probably numbered. And then there are those who, for perfectly good reasons, are sensitive about doing some of their shopping in person, perhaps because of their weight or sexual tastes. Consumers should not have to pay the price of surveillance every time they want to buy a pair of pants or a sex toy. Also, in-store facial surveillance will make it increasingly difficult to shop there anonymously. In fact, online shopping could in theory increase privacy, since it means that shoppers can lie low (Moor 1990). Again, if I am justified in speaking softly or in donning disguises, then I am justified in obfuscating my online presence. In fact, the more people who successfully obfuscate, the less motivated will retailers be to collect the information and the less valuable it will be as a commodity.

Another objection, derived from Richard Posner, might be that obfuscation impairs the kind of market efficiency that big data has promoted (see Posner 1978a, b, 1979, 1981a, b). The vast majority of us have an interest in greater transparency in the marketplace, and obfuscation undermines this. Surely, it is better for both parties if marketers promote their products only to those who are interested in buying them rather than wasting their efforts on consumers who could not care

less. The savings from targeted marketing and advertising can be passed along to consumers. Also, perhaps privacy has greater costs than I have acknowledged. The coin of the realm in social relations is reputation: how people regard another as a friend, potential mate, colleague, or business associate. Since reputation is people's best social asset, they are strongly motivated to hide what they think will harm it. People "sell" themselves as a merchant sells products or services. People often plead for privacy to give themselves an advantage in their dealings with others, and sometimes this advantage is unfair. Take credit scores. People will want to hide a bad one both because it limits their opportunities in the marketplace and because it harms their reputation generally. Insisting on a "right" to privacy in this context creates inefficiencies both in the market for loans and in the social "market" for mates, friends, and associates. Privacy, as promoted by obfuscation-based anonymity, chokes the supply of valuable information. Banks would prefer to lend only to those who will make good on their loans. Who can blame them? We saw above that credit scores speak volumes about insurance risk: the lower one's score, the more likely one is to make a claim (Rule 2007). If credit scores, along with much other information about people's habits, are a reliable way to evaluate risk, why should insurers not have this information? More generally, whether people pay their bills on time or at all says a lot about their characters. People with low credit ratings often get a discount on the social and financial price of being deadbeats. Why are the rest of us stuck with the tab? Also, those with good credit scores or who match other criteria for upstandingness will benefit from the unimpeded flow of financial information in all kinds of ways, starting with lower interest rates and insurance premiums. Here, as elsewhere, privacy shifts the social and economic costs in the wrong direction. Instead, we should let the market, including the social market for personal information, sort things out.

Again, I acknowledge the costs of my proposal but maintain that the current arrangement works to the extreme disadvantage of data subjects for the reasons that I have canvassed above. The market and the government alike have failed to protect us from this serious threat, with no serious prospects that either will offer succor in the future. And even if we assume that the market will eventually offer genuine protection of online privacy, this protection will come at a price, turning privacy into a luxury. Under the circumstances, those who are concerned about their vanishing privacy have no serious alternatives to anonymity through obfuscation.

20.5 The Tyranny of Small Decisions

Alfred Kahn (1966) persuasively argues that individuals routinely make decisions that are entirely rational from their point of view but which can lead to sub-optimal results in the aggregate. In other words, virtually no one would have chosen the outcome of these small decisions, each of which was nevertheless in the interest of the individuals who made them. His argument has direct relevance for the loss of privacy in the light of advances in information technology that discrete decisions

have encouraged (see Rule 2007). Kahn's own example involves the elimination of passenger trains from his relatively isolated town in the 1960s. Virtually every time a person had to leave town it was rational for him to choose to drive or fly, given the inconvenience and greater expense of rail. Unfortunately, the outcome of thousands of such choices was the elimination of train service, a result that almost no one would have chosen beforehand. After all, the train was the most reliable way to travel in foul weather. The trouble is that once trains are gone it becomes colossally expensive to bring them back, since investment will be shifted from them to roads and airports.

A similar tyranny could be promoting the decline of brick-and-mortar shopping. Online shopping offers many benefits over offline. The former wins in terms of convenience. It also tends to be cheaper, since it makes comparison shopping much easier, although this benefit could be short-lived, as retailers, through profiling, get better at figuring out how much online customers are willing to shell out. The trend obviously bodes ill for conventional stores. The more people shop online, the fewer stores there will be. Assume that every time that a person decides to shop online as opposed to offline she is saving either money or time. Second, assume that everyone is striving to maximize their own welfare. Then it will always be irrational for people to shop offline instead. The purchasing power of a single consumer cannot affect the market at large (Kahn 1966). So it will not be rational for a person to consider the possible negative effects of her shopping decisions on traditional retail. Yet the outcome in which there is far less brick-and-mortar than at the outset might be something that no one would have chosen had they been given the choice from the start, both because once lively neighborhoods are now forlorn or derelict and because of the loss of privacy. People might well have been willing to pay more and give up some convenience had they anticipated the upshot of their collective decisions. In such a case, brick and mortar is the victim of the tyranny of small decisions, decisions which are rational for individuals to make but whose cumulative effect practically no one wants and would have rejected outright had the result been presented "for their explicit consideration" (Kahn 1966, p. 24). Part of the problem might be that the full costs, economic and social, are not included in the price of online shopping. Had they been, it would not have been rational for people to shift their buying habits as they have and will no doubt continue to do. Kahn makes a similar point about driving. If drivers were forced to pay the full price of driving, including pollution, noise, danger to others, sprawl, and climate change, there would be far less driving and probably more support for trains. Currently retailers and giant datamongers like Choicepoint and Axciom are not paying full freight. If they were, then data brokerage and the resulting surveillance would be less profitable and online shopping would not be so cheap. Finally, as with trains, so with online shopping, the tyranny of small decisions means the nearly irreversible elimination of alternatives. Even if we wanted to go back to the old days of comfortable and efficient trains, the investment needed to do so would be huge, even prohibitive (Kahn 1966). Conventional retail might be headed the same way and with it a good measure of our privacy in public.

20.6 Autonomous Technology?

Some might despair, opining that information technology is a quasi-Hegelian force with an unslakable thirst for ever more information, operating according to an "internal dynamic" (Winner 1980, p. 122). Such a view has defenders. The notion seems to be that technology's march is independent of human will or purposes, that it is "self directing," proceeding according to its own "necessity," "laws," "logic" or "imperatives" (Ellul 1989, p. 135; Winner 1977, pp. 13 and 15; Rule 2007, pp. 18–19; 33; 160).[1] Langdon Winner documents how thinkers from John Kenneth Galbraith to Heidegger have thought of technology as something irresistible, pursuing its own course with a "self-propelling, self-sustaining, ineluctable flow" (Winner 1977, p. 46). Says Heidegger: "Technological advance will move faster and faster and can never be stopped. In all areas of his existence, man will be encircled ever more tightly by the forces of technology. These forces ... have moved long since beyond his will and have outgrown his capacity for decision" (quoted in Winner 1977, p. 14). Technology generally, and information technology in particular, mows down values like privacy and autonomy as it advances. Says Jacques Ellul, a leading exponent of the view: "There can be no human autonomy in the face of technical autonomy" (quoted in Winner 1977, p. 16). So we might as well step aside.

It is true that technology has developed in ways that no one, even self-styled futurists, could have predicted. No one in 1935 saw the digital revolution coming; no one in 1985 envisaged just how central the internet would be today. As mentioned above, not even big data holders could have anticipated the remarkable discoveries that data analysis has yielded or just how cheap storage would become. Nevertheless, *we* create and refine technology, and *we* need to make it do our bidding. We should resist the canards from the titans of big data about changing norms, long-lost privacy, or the unstoppable thrust of technology. There is no reason to suppose that the information and power asymmetries between us and big data have anything to do with the intrinsic nature of technology. A simpler explanation is that privacy is under siege from retailers and data brokers, not because technology is indomitable but, as James Rule puts it, because "one group is simply stronger and better organized than the other" (Rule 2007, p. 20). Moreover, as pointed out above, technology can actually promote privacy in a wide range of cases.

Obviously, much commercial surveillance is driven by the huge online advertising market. Advertising is part and parcel of a market economy, and it makes sense that merchants and advertisers are eager to match products and services to those who want them at prices they are willing to pay. John Wannamaker evidently quipped that he knew that half of what he spent on advertising was wasted. The trouble was that he didn't know which half (Schneier 2015). From a business point of view this is deeply unsatisfactory. It means that much of what companies spend on traditional advertising will be good money after bad. However, we need to weigh this cost against the havoc wrought by big data to privacy and autonomy. It is far from clear

[1] Winner and especially Rule are critical of the notion of autonomous technology.

that turning the internet into a cash cow is worth the thoroughgoing surveillance that results. And unless we take action by, say, criminalizing nearly all secondary uses of personal information, things will go from bad to worse for privacy as the techniques of data analysis become more sophisticated. Meanwhile obfuscate.

20.7 Conclusion

I have tried to defend strict online anonymity through obfuscation as a tool for protecting privacy and thus autonomy. I acknowledge that there are potential costs to anonymous web searching. I further acknowledge the tradeoff between the anonymization of data and its utility. The more anonymized data is, the less useful it is (Ohm 2010). However, I have tried to put the burden on those who would defend the current regime of surveillance. It seems to be dogma among commercial data holders and their partisans that the internet exists chiefly as a profit-maximizing domain and that anything that can be "monetized" there should be to the fullest extent possible (Nissenbaum 2011). I reject this assumption in the name of privacy and autonomy. To the extent that the internet exists to maximize the profits of big data and their clients, privacy and autonomy lose. In such a regime, online privacy will become a luxury and thus out of reach of the poor (Angwin and Steel 2011; Tavani and Moor 2001).

Given the role that privacy plays in ensuring autonomy, I would prefer to see the former treated as a human right, like other rights regarded as essential for autonomy and democracy: expression, access to information, and assembly. This would mean that privacy is no more a prerogative of the rich than these other fundamental legal rights. An important question is, can we continue to enjoy the considerable benefits of big data without destroying privacy in the process? Unfortunately, those who predict or have pronounced the end of privacy might be right. However, if they are, it will not be because privacy's demise was inevitable but because those inimical to it hold all the cards. My concern is that we will "accept" the current trends in surveillance not because they are on balance beneficial or ineluctable but because the pro-surveillance forces are more powerful and better organized than the subjects of surveillance. Consider climate change, perhaps the most serious challenge now facing humanity. My guess, and I hope I am wrong, is that we will not get the kind of international cooperation to stop the worst of the damage. Between inertia and the powerful forces that have an interest in continuing a petroleum-based economy, I do not see much hope for a solution. Powerful and influential forces too are aligned behind the current big data regime. Maybe it's time to steel ourselves for the end of privacy. I hope not. However, as long as the obfuscation arms race is worth running, perhaps we can find some refuge online.[2]

[2]I would like to thank Jane Carter, Don Fallis, and Catherine Womack for their comments. Also, I presented earlier versions of this paper at the 2016 Information Ethics Roundtable, held at the University of Arizona, Tucson, and at the 2016 annual meeting of the International Association of Computing and Philosophy, held at the University of Ferrara, Ferrara, Italy. I would like to thank participants for their feedback.

References

Angwin, J. 2014. *Dragnet nation: A quest for privacy, security, and freedom in a world of relentless surveillance*. New York: Times Books, Henry Holt and Company.

Angwin, J., and E. Steel. 2011, February 28. What they know: A *wall street journal* investigation: Web's hot new commodity: Privacy. *The Wall Street Journal*, p. A1.

Aquisti, A. 2014. *The economics and behavioral economics of privacy*. In Lane 2014.

Barbaro, M., and T. Zeller. 2006, August 9. A face is exposed for AOL searcher no. 4417749. *The New York Times*, A1.

Barocas, S., and H. Nissenbaum. 2014. Big data's end run around anonymity and consent. In *Privacy, big data, and the public good: Frameworks for engagement*, ed. J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, 44–75. New York: Cambridge University Press.

Brunton, F., and H. Nissenbaum. 2011. Vernacular resistance to data collection and analysis: A political theory of obfuscation. *First Monday* 16 (5).

———. 2013. Political and ethical perspectives on data obfuscation. In *Privacy, due process and the computational turn: The philosophy of law meets the philosophy of technology*, ed. M. Hildebrandt and K. De Vries, 164–188. New York: Routledge.

Cohen, J. 2000. Examined lives: Information privacy and the subject as object. *Stanford Law Review* 52 (5): 1373–1438.

Duhigg, C. 2009, May 17. What does your credit-card company know about you? *New York Times Magazine*, 40–45.

Ellul, J. 1989. *What I believe*. Trans by G. Bromiley. Grand Rapids: Eerdmans.

Gandy, O. 1993. *The panoptic sort: A political economy of personal information*. Boulder: The Westview Press.

Howe, D., and H. Nissenbaum. 2009. Trackmenot: Resisting surveillance in web search. In *Lessons from the identity trail: Anonymity, privacy, and identity in a networked society*, ed. I. Kerr, C. Lucock, and V. Steeves, 418–436. Oxford: Oxford University Press.

Kahn, A. 1966. The tyranny of small decisions: Market failures, imperfections, and the limits of economics. *Kyklos* 19 (1): 23–47.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big data: A revolution that will transform how we live, work, and think*. Dolan/Houghton Mifflin Harcourt: Boston and New York.

Moor, J. 1990. The ethics of privacy protection. *Library Trends* 39 (1–2): 69–82.

Nissenbaum, H. 1998. Protecting privacy in the information age: The problem of privacy in public. *Law and Philosophy* 17 (5–6): 559–596.

———. 1999. The meaning of anonymity in an information age. *The Information Society* 15 (2): 141–144.

———. 2010. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford: Stanford Law.

———. 2011. A contextual approach to privacy online. *Daedalus* 140 (4): 32–48.

Oberhauser, K. 2011. Monarch butterfly. In *Environmental encyclopedia*, vol. 2, 4th ed., 1091–1093. Gale: Detroit.

Ohm, P. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57 (6): 1701–1777.

———. 2014. Changing the rules: General principles for data use and analysis. In *Privacy, big data, and the public good: Frameworks for engagement*, ed. J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, 96–111. New York: Cambridge University Press.

Posner, Richard. 1978a. The economic theory of privacy. *Regulation* 19 (2): 19–26.

———. 1978b. The right to privacy. *Georgia Law Review* 12 (3): 393–422.

———. 1979. Privacy, secrecy, and reputation. *Buffalo Law Review* 28 (1): 1–55.

———. 1981a. *The economics of justice*. Cambridge, MA: Harvard University. Press.

———. 1981b. The economics of privacy. *The American Economic Review* 71 (2): 405–409.

Rule, J. 2007. *Privacy in peril*. New York: Oxford University Press.

Schneier, B. 2015. *Data and Goliath: The hidden battles to capture your data and control your world*. New York: W.W. Norton.
Singer, N. 2013, September 1. A data broker offers a peek behind the curtain. *New York Times*, BU1.
Steel, E., and J. Angwin. 2010, August 4. On the web's cutting edge, anonymity in name only. *The Wall Street Journal*, A1.
Tavani, H., and J. Moor. 2001. Privacy protection, control of information, and privacy-enhancing technologies. *Computers and Society* 31 (1): 6–11.
Throsby, D. 2015. Trust funds. *TLS* 5861: 28.
Walker, J. 2013, December 17. Data mining to recruit sick people. *Wall Street Journal.*
Wallace, K. 1999. Anonymity. *Ethics and Information Technology* 1 (1): 23–25.
Winner, L. 1977. *Autonomous technology: Technics-out-of-control as a theme in political thought*. Cambridge, MA: MIT Press.
Winner, L. 1980. *Do artifacts have politics?* 109 (1): 121–136.
Zimmer, M. 2010. "But the data is already public:" On the ethics of research in Facebook. *Ethics and Information Technology* 12: 313–325.

Chapter 21
Safety and Security in the Digital Age. Trust, Algorithms, Standards, and Risks

Massimo Durante

Abstract Security is a crucial issue of our society, which is accordingly defined as a risk society. However, in a complex risk society, citizens cannot tackle and manage the issue of risk by themselves. The risk is therefore more and more delegated to processes and mechanisms that take care of risk management. Today, the risk against which society claims to be immunized-increasingly mediated by technologies and less and less politically legitimized-reemerges with new forms of fiduciary management, raising the possibility of weakening rights and diluting political responsibility.

Keywords Trust · Security · Artificial agents · Delegated authority · Risk-management · Fiduciary relationships · Information

21.1 The Relation Between Trust and Security

In our contemporary information societies, there is a great, resurgent emphasis placed on the issue of security from many perspectives (i.e., from legal, economic, social, ethical, technical and, notably, political standpoints). There is a widespread, almost compulsive, demand for more and more security. We aspire to secure almost every aspect of our daily life as we strive to better ensure private and public decisions. However, we can neither handle all daily requirements of this quest for increasing security nor be involved in every private or public decision. Security requires effort, time, and means we do not necessarily have on our own. Thus, we need to rely on something else (i.e., human or artificial resources, means, and devices). Alternatively, we need to delegate this task to someone else (i.e., human or artificial agents [Durante, 2010]). In our digital age, for instance, we need more and

M. Durante (✉)
Dipartimento di Giurisprudenz, Univeristà di Torino, Turin, Italy
e-mail: massimo.durante@unito.it

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_21

more to deal with the social (ethical, legal, economic, and political) impact of the delegation of decisions to automated systems and autonomous artificial agents. For better or worse, this process of delegating decisions to both automated systems and autonomous artificial agents is already transforming the environment of peoples' interactions and their daily lives (Pagallo and Durante, 2016). Yet delegation is not socially acceptable or enforceable if it is not supported by (a more or less distributed) trust. Either at an individual or at a systemic level, people need to trust, at least to some extent, their delegates or proxies (Luciano Floridi has, indeed, recently qualified our own culture as a *culture of proxies* [2014, 58][1]). The more security is systematically delegated to entrusted agents or devices, the more security presupposes (actual or perceived; individual or systemic) trust. This is the reason why we intend to examine the relationship between security and trust, which has been mainly overlooked, if not disregarded, in the overall tradition of political and legal studies.

Let us remark at the outset on the theoretical relevance and framework of such a crucial relationship both in modernity and in our contemporary technological, and mainly digital, age.

In modernity, at least since Hobbes, the father of modern political and legal thought, security is at the center of any political and legal project in its *negative* dimension: the want of security. A person's life is always menaced by the threat of some evil ("the war of all against all") and therewith their life projects: "In such condition, there is *no* place for industry; because the fruit thereof is *un*certain: and consequently *no* culture of the earth; *no* navigation, *nor* use of the commodities that may be imported by sea; *no* commodious building; *no* instruments of moving, and removing, such things as require much force; *no* knowledge of the face of the earth; *no* account of time; *no* arts; *no* letters; *no* society; and which is worst of all, continual fear, and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short" (Hobbes, 1651; 1991: XIII, 89; emphasis mine). There is no room in the present context to look at Hobbes' strategic and systemic use of *negative* terms, but suffice it here to say that he is the creator and harbinger of the negative condition of *insecurity* (uncertainty, fear and so forth) as the foundation of the civil society from a political and a legal standpoint.

In this *negative* perspective, the want of security becomes the basis of political and legal authority and the fundamental lever that allows a government to gain greater powers of direction and control with the aim of immunizing society against the risks that seem to threaten its integrity. Security is thus understood and interpreted as the central issue of the political (legal, economic, social) project of modernity, according to which it is also seen and presented as a critical process of risk-management. It is exactly the aforementioned negative dimension that turns

[1] "And since in the infosphere we, as users, are increasingly invited, if not forced, to rely on indicators rather than actual references–we cannot try all the restaurants in town, the references, so we trust online recommendations, the indicators of quality–we share and promote a *culture of proxies*. LinkedIn profiles stand for individuals, the number of linked pages stand for relevance and importance, 'likes' are a proxy for pleasant, TripAdvisor becomes a guide to leisure".

security into a process of risk-management, along the Hobbesian line of reasoning: there is *no* place for all of this (i.e., 'culture', 'navigation', 'building', 'instruments', 'knowledge' and so forth, which are then fatally put at risk), if there were no possibility for the political authority to grant security to all. And this is so in two essential respects. Security is hence understood as (1) a fundamental good (concerning the integrity of individuals' life and of their life projects), related to each individual, and as (2) a condition for the enjoyment of goods, namely, the implementation of life projects (concerning the community in its entirety). In the first case, security is a goal of political (legal, economic, social) life; in the second case, it is a condition of possibility of political (legal, economic, social) life.

The distinction between the goal and the condition of possibility of political life can be traced back to and expounded through the distinction between the two connected concepts of *safety* and *security. Safety* is mainly aimed to ensure the integrity of life against the threat of imminent dangers. It has a temporal dimension associated with immediate relationships (*e.g*. the violent dimension of time that, according to Locke [1690; 1998, 3.19],[2] does not leave us the time to delegate our decisions to the authority of a third person). *Security* is mainly aimed at the protection of the conditions for the enjoyment of goods against the threat of dangers that may be subject of anticipation and calculation. It has an inter-temporal connotation that is mainly part of mediated relationships (*e.g*. the dimension of time underlying the instrumental rationality). From the modern to the contemporary ages, security has an increasingly central role in politics. Whether understood in terms of *safety* (protecting the integrity of individuals' lives) or *security* (protecting the integrity of shared life projects), security is a key political lever, allowing a government to adopt penetrating powers of direction and control by promising to immunize society against the risks that threaten its integrity. However, this immunization is always relative to specific threats and risks, not only because one cannot imagine an existence placed away from any threat or risk (Beck 1992), but also for a more basic reason, which has been remarked several times in the relevant political literature, but not always brought to light and investigated systematically. This reason is the complex and problematic relationship security has with trust.

In fact, as already pointed out, even in the simplest societies, security is neither managed nor granted on citizens' own. On the contrary, it is rather delegated through fiduciary forms that tend to institutionalize the risk-management (Giddens 1990, 26).[3] Thus security needs delegation and hence trust. However, where there is trust, there is risk (Luhmann 1979). Here lies our basic idea we would like to explore and argue for in the present paper. Today, the risk against which society claims to

[2] "The law, which was made for my preservation, where it cannot interpose to secure my life from present force, which, if lost, is capable of no reparation, permits me my own defense, and the right of war, at liberty to kill the aggressor, because the aggressor allows not time to appeal to our common judge, nor the decision of the law, for remedy in a case where the mischief may be irreparable".

[3] "All disembedding mechanisms, both symbolic tokens and expert systems, depend upon *trust*. Trust is therefore involved in a fundamental way with the institutions of modernity".

be immunized emerges again with the new, increasingly mediated by technologies and less and less politically legitimized, forms of fiduciary risk-management, where the possibility of a weakening of rights and of political responsibility resides. This requires studying both the ways in which the fiduciary forms of risk-management have been structured and institutionalized throughout modernity and how these fiduciary forms presently depend on the ongoing process of digitization of current societies and politics. In modernity, the fiduciary forms of risk-management have taken place, mainly in three ways, through:

1. The creation of fiduciary relationships between actors [*e.g.* the social contract];
2. The delegation of risk-management to legal and technological devices [e.g. identity documents; data collection, etc.]; and,
3. The delegation of risk-management to mixed organizations (made by actors and devices [e.g. tax authorities]).

At present, we observe a twofold evolution of the fiduciary forms of risk-management, which is tied up with the technological, digital evolution of our societies and politics. On the one hand, the fiduciary forms of risk-management are progressively incorporated in automated technological devices, according to the processes of algorithms escaping, at least partly, from direct human control. This raises an issue of a growing political importance: i.e. the government of algorithms (Barocas et al. 2013), whose automated and neutral functioning seems to rule out some forms of human understanding and control. On the other hand, the fiduciary forms of risk-management are deployed today in a context in which the concept of politics is changed. Politics is no longer understood only as a form of control over a territory or a way to take collective decisions, but mostly as a form of control over the "public mind" (Castells 2009, 56)[4] or, to put it otherwise, over the life cycle of information (Floridi 2010). Information is increasingly an essential resource by which we represent the world and we make collective decisions (Benkler 2006, 1).[5] Informational resources build the standards by which we mediate our relationship with reality. This raises a further and critical aspect of contemporary forms of power: that of the government of standards (for the relevance of standards in contemporary society see Busch 2011, 13).[6] Algorithms (with their automated and impersonal efficiency) and standards (with their capacity to mediate the normative

[4] "The *public mind* – that is, the set of values and interpretative frames that have large exposure in society - is ultimately what influences collective behaviour".

[5] "Information, knowledge, and culture are central to human freedom and human development. How they are produced and exchanged in our society critically affects the way we see the state of the world as it is and might be; who decides these questions; and how we, as societies and polities, come to understand what can and ought to be done. For more than 150 years, modern complex democracies have depended in large measure on an industrial information economy for these basic functions".

[6] "Standards are means by which we construct realities. They are means of partially ordering people and things so as to produce outcomes desired by someone. As such, they are part of the technical, political, social, economic, and ethical infrastructure that constitutes human society".

understanding of reality) will be at the core of the political debate about trust and security in the years to come. In both cases, these modes of government—and, in particular, the delegation of the forms of risk management—depend on the pervasive technological dimension of the information society. Today, politics is actually intertwined with the growing dependence of modern societies on information and communication technologies and on the greater convergence between offline and online realities. Since the map of our dependencies is the map of our vulnerabilities, this requires us to examine more closely the relation between trust and security in the digital age from the perspective just outlined.

Against this general theoretical framework, our paper has to face at least three main questions: § 2. What is the relationship between trust and security such that they mutually depend on each other in the sense that they can weaken and reinforce each other? § 3. How does the risk-management is progressively incorporated into technologically automated processes and devices? § 4. What is the increasingly key relationship between information, trust and security or, to put it otherwise, the growing relevance of the informational understanding of the relationship between trust and security? On the basis of those analyses, we will draw some final remarks in the last paragraph of the present paper.

21.2 The Mutual Relationship Between Trust and Security

The gradual shift from risk-management mediated by human relationships to technologically mediated management can display a mutual relationship between trust and security in the sense that they can weaken and reinforce each other.

A. Consider, first, the case of *mutual reinforcement,* in which the two elements of the relationship reinforce each other:

   1. The sense of security is reinforced by the sense of trust. This happens, for example, in the case in which we are confident in those to whom the management of security is entrusted. The growing lack of trust in political leaders has undermined this relationship of trust and with it the delegation of risk-management, which is therefore directed towards forms of incorporation of security in mixed organizations or rather in automated technological devices, which tend to turn the trust into reliance and to make such management more predictable and controllable, but also more sterile, impersonal and anonymous.
   2. The sense of trust is reinforced by the sense of security, because the sense of security allows us to expose ourselves to a greater degree of risk and to develop relations of trust with some stability. This was true for interpersonal trust relationships but remains true in case of relations based on automated technological devices, the (relatively) predictable and standardized operation of which is characterized, at least in principle, by a high degree of stability. Furthermore, the pervasiveness of the incorporation of trust into automated

devices and processes, which are more and more widespread, tends to create an overall sense of confidence in all areas of application. In this sense, in the technological age, trust is less and less attributed with regards to specific cases and relationships (outside of a detailed technical ability to have control over the delegation process). On the contrary, it is attributed on a larger scale with regards to more generic (technologically mediated) contexts and platforms. In other terms, the issue of risk management is no longer only a question of relational trust, expressed with regards to specific actors. It is a matter of systemic trust, expressed in relation to the system those actors are part of (in the political perspective see Durante 2012, 8).

B. Let us consider, then, the opposite case. This is the case of *mutual weakening,* in which the two elements of the relationship weaken each other:

1. Trust is more easily betrayed where we feel more secure, that is, in the sphere of our securities: in this case, the sense of security makes us more vulnerable. This has always been true in interpersonal relationships (as well observed by Nissenbaum 2004), but remains equally true in relationships technologically mediated, where reliance on automated and standardized processes and devices seems to guarantee more efficient and effective operations, at the cost of a real and shared capacity to have control over such operations (e.g., once we delegate the control over the circulation of persons to the guaranties provided for by a biometric passport [Lyon 2009], this seems also to relieve us from the need of any further investigation).
2. Security is more easily jeopardized where we have trust (or overconfidence) on who or what is responsible for the risk-management. This aspect is perhaps the thorniest of the case at hand. The incorporation of trust in technologically mediated processes and devices tends to fatefully elide needed and crucial space for discussion, interpretation and public discussion, in which we might scrutinize and evaluate the technologically mediated forms of risk-management. The lack of a sphere of public reflexivity and judgment tends to turn political action into a form of more or less sophisticated administration by weakening the dimension of political responsibility and accountability. Therefore, this aspect raises the need to look more closely to the incorporation of risk-management into standardized and automated processes and devices.

21.3 The Incorporation of Risk-Management into Technological Processes and Devices

The incorporation of risk-management in mixed forms and largely in standardized and automated technological processes and devices is a feature of our technological, digital age (Pagallo 2013). This incorporation has certain undeniable positive

aspects arising from the process of standardization and automation in terms of efficiency, predictability, consistency and reliability of processes and technological devices. At the same time, it raises some problematic aspects in terms of its political and democratic legitimacy (Pagallo and Durante 2016). Think of the case, already mentioned, of the biometric passport. Generally understood, a passport is a mixed form of incorporation of trust: it enables certain operators to verify that beholders of a passport have some qualities by means of a relatively simple control mechanism. The biometric passport involves the incorporation of an additional standard that may end up making, in certain contexts, the ordinary passport obsolete (as if the beholder of an ordinary passport should still justify its failure to comply with a given supplementary standard, which characterizes the assessment of the new individual 'identity' [Bauman and Lyon 2012, 126][7]). In this case, the standard built into the device (e.g. the biometric passport) has a twofold role (as it happens for many standards, as properly remarked by Lawrence Busch 2011, 18–20): it explicitly assures uniformity but it is also the implicit supporter of some qualities, which have a normative role (they allow, or prevent, people to cross a border). This is just an example. In what follows, we have to expand our analysis. Let us hence point out three main problematic aspects of the process of the technological incorporation of risk-management.

21.3.1 The Invisibility of Computer-Based Technologies

First, it should be remembered that one of the salient features of computer-based technologies is their relative *invisibility*: "most of the time and under many conditions the computer operations are invisible" (Moor 1985). On the one hand, the invisibility of computer operations is a factor of extreme efficiency, because it does not require the constant user's intervention or control; on the other hand, it is problematic exactly because of the fact that this invisibility makes us more vulnerable to risks and uncertainties that the user is not always in a position to perceive and master. This causes problems resulting from the different ways in which the factor of invisibility is configured. There are at least three diverse types of problems: (1) problems brought about by *invisible abuses*; (2) problems brought about by *invisible program values;* and finally (3) problems brought about by the *invisible computing complexity.*

[7] "In numerous surveillance situations, bodies are reduced to data, perhaps most obviously through the use of biometrics at borders. Yet in this paradigmatic case, the end in view is to verify the identity of the body, indeed, of the person, to permit them to cross the border (or not). One cannot but conclude that information *about* that body it is treated as if it were conclusive in determining the *identity* of the person. [. . .] In condensed form, this is the story of how disembodied information ends up critically affecting the life chances of flesh-and-blood migrants, asylum seekers and the like".

In the first group, we can include those cases where the intentional use of invisible operations aims at achieving a potentially wrongful act: cases of violation of privacy or data protection, of unauthorized surveillance and so forth. Within the second group, we can mention those cases where, in a given program, invisible values are implemented (Brey 2010), so that concrete choices are filled with hidden value judgments which are not visible in the results of an operation made by means of a computer program. Sometimes the implementation of values is invisible to the programmer as well, who is not always in the position to calculate the unintended consequences arising from the choices made during programming. In the third group, we can bring those cases in which the enormous complexity of calculation (which is invisible in the sense that its verification requires a highly complex and uneconomical process) is susceptible to determining a series of consequences not entirely predictable and calculable: it is the case in which an activity performed by computer is based on the trade-off between the reliability of transactions based on a calculation that is likely to be more efficient from a probabilistic standpoint than the human calculation and the impossibility of direct control over the all stages and consequences of the performed activity.

In this perspective, the incorporation of security management in the technological processes and devices can also: (1) give rise to *invisible abuses,* as in cases where the monitoring of persons does not result in a direct intrusion into the privacy with the uncontrolled acquisition of a huge amount of personal data; (2) convey values, in a more or less surreptitious way, in the *design* of technological processes and devices, resulting in hidden forms of paternalism, which limit the space of personal choices, shape social behaviours and have discriminatory consequences; and, (3) have *socially relevant consequences*, in which it is difficult to figure out and trace the terms of (political, legal, moral etc.) responsibility.

21.3.2 The Absence of a Mediation in Terms of Understanding, Debate and Public Scrutiny

The mechanism of incorporation of risk-management into technological processes and devices tends to cancel out, or at least limit the sphere of understanding, discussion and public scrutiny that serves as an interface between the institutionalization of forms of security and their trustworthy application (Zittrain 2010). This interface not only plays a crucial role in the process of political legitimacy and democratic management of security but may also assure or restore some degrees of social acceptance, notably where such a management gives rise to controversial cases in regulatory contexts that are not characterized by a high degree of agreement and shared values, norms and principles (which is often the case in our pluralistic societies). In fact, the higher the degree of social cohesion (about values, norms and principles) in a given regulatory context, the greater the degree of social acceptability of the risks inherent in the delegation of security management (Pagallo and

Durante 2016). In this perspective, the standardized and automated characters of risk management tend to limit the public sphere of interpretation, debate and scrutiny, which is essential to the formation of a sufficiently informed, aware and responsible level of social acceptance. There is also a significant further consequence. The ongoing removal of a sphere of mediation between the request of fiduciary forms of risk management and their standardized and automated application tends to turn *security* issues into *safety* issues (understood and represented as more urgent and impending). This allows the governors to ask for further power of direction and control as well as for reinforcing the technological dimension of risk management, with a paradoxical result. In fact, it has not been sufficiently remarked that, in the age of widespread security, claiming a permanent threat of sudden and unpredictable dangers is the basis of the incorporation of risk-management into increasingly standardized and automated technological processes and devices. For example, the pretended need to prevent crimes, cyber-attacks and other forms of terrorism justified the request to implement automatic and widespread filtering systems for the collection of personal data and information, which gave rise to the well-known scandal of North American Prism Agency for national security (NSA) and the so-called British GCHQ files, which was severely judged, in Europe, by the European Court of Justice for being indiscriminate and illegitimate (C-360/10, §§ 50) (see on this point, Pagallo 2014, 174–183).

21.3.3 The Empowering Nature of the Enabling Technological Processes and Devices

Enabling technologies create possibilities that change the environment in which people act and are likely as well to make obsolete existing law, to modify social and political reality, and to affect the distribution of power in the society. In this sense, technology empowers some and dis-empowers some others. It is because of this redistribution of power that technology affects politics, law and the economy, and forces, from time to time, rethinking the conditions and manner in which power is created or reallocated (Durante 2007, 284–289). The redistribution of power, engendered by technological evolution, is neither necessarily politically justified (through some form of democratic approval) nor always legally ratified (*ex ante* or *ex post* by a parliament or a court). In the digital era, law is no longer the only normative system regulating the distribution and redistribution of power. Law is a normative system in competition with other regulatory systems (code, social norms, economics, architecture, standards, algorithms and so forth), as remarked in a different context by Lawrence Lessig (1999, 2006). The technological redistribution of power is therefore susceptible to change, in a more or less perceptible way, with respect to some institutional arrangements and the allocation of power among actors: this also involves the incorporation of security management in technological processes and devices, which hence may play a normative and regulatory role

by incorporating norms and values by design. Needless to say, this raises crucial legal and political problems when the incorporation of security management into technological processes and devices alters substantially or formally the democratic allocation of powers between political or social actors. As already remarked in the previous sub-paragraph, the political demand for more incisive governmental powers for monitoring and collecting data has recently exploded into well-known scandals as a direct result of a widespread perception of trust betrayal. To some extent, the latest politics of security has been, above all, a politics of scandals. The series of scandals that have accompanied the politics of security not only indicate the redistribution of power that technology has enabled (often ahead of any real political or legal legitimization) but also the asymmetric information surrounding this redistribution of power. From this perspective, the asymmetry of information is one of the constitutive feature of political power in our current information societies. We must, therefore, focus our attention, in what follows, on the critical issue concerning the relationship between information, trust and security.

21.4 The Relationship Between Information, Trust and Security

The information and communications technologies are reontologising the reality: they build a world made by information: a real "informational habitat" (Kallinikos 2006, 2011) or, to put it in Luciano Floridi's terms, an "infosphere" (Floridi 2010) that exceeds the dichotomy between the offline and online realities (Floridi 2013, 2015). According to this view, information forms the basis on which we represent and interpret the world. We make decisions and act in accordance with the expectations, knowledge and culture that we elaborate through the informational resources to which we have access. Against this backdrop, politics can no longer be understood solely as the sphere of collective sovereign decisions (normatively) or the legitimate monopoly of force and control over a territory (descriptively). It should rather be conceived, more modernly, as a form of management and control over the life cycle of information (that implies the production, circulation, storage, distribution, redistribution and deletion of information: Floridi 2012), which influences and conditions the formation of the public mind, political consent, the decision-making process, and the legitimacy of political power. This concept of politics, elaborated in the framework of technologically advanced societies, also affects the security policies and the incorporation of security into ever more technologically mediated, fiduciary forms of risk management. This requires us to consider more closely the relationship that security and trust have with information. Needless to say, our remarks are only one possible sketch of a more intricate and manifold relationship.

The control over information is vital to any project or security policy. The security of x is greater, the greater the information that x holds and is able to

model accessing. That is, the security of *x* may depend not only on the amount of information that *x* holds but above all on x's ability to model the access that others have to such information. Someone's security may actually depend on the information she prevents the others from accessing (*e.g.* where one has hidden the stolen goods) or, on the contrary, on providing others with information (*e.g.* what is her blood group). Security depends on control over the information (or rather over the life cycle of the information). Their control is a way to manage security. However, as some recent experiences may teach us in a tragic way (think of many cases of terroristic attacks), the control over the information (as a prerequisite of security) depends in turn on the ability of an agent to have access to *relevant* and *reliable* information out of a big collection of data. This raises fundamental question about cognitive attention and epistemic trust in information societies. When societies are characterized by increasing informational overload (i.e. when information is no longer a scarce resource), it becomes crucial how we choose which information to attend (i.e. the relevance of information); and how we choose which information is reliable (i.e. the trustworthiness of information). Cognitive attention and epistemic trust are also deeply intertwined with relational trust from a political standpoint, since we are gradually but increasingly called upon to trust who is in charge to designate or decide which information is relevant and reliable.

Furthermore, both relational and epistemic trust involve a key relationship to information. Trust is ruled out by two opposite situations: the completeness of information (certainty does not really require any act of trust); the radical incompleteness of information (ignorance requires only an act of blind trust). Trust has always a necessary and essential relationship with incomplete information: that is, trust depends on the incompleteness of information and, in that respect, it is one of the best way to deal with and manage their incompleteness. From this point of view, trust and security have a similar attitude towards information. Trust is associated to a certain degree of information: either too much or too little information tends to exclude the circumstances in which it matters to trust. Security seems to have an even more scalable relationship with information. More information (and wider control over information) produces a greater (sense of) security, with a significant limitation, however: there is a point at which the flow of information (i.e., information overload) makes it difficult to have access to the relevant and trustworthy information that is indispensable for the efficient implementation of any security policy.

All these aspects raise a number of questions that are meant to frame the policies of information in our digital era, as it has been pointed out recently by Luciano Floridi in many workshops: (1) who has access to what information (i.e. the issue of the availability of information)? (2) Who has access to the relevant and reliable information (i.e. the issue of control over filters of information)? (3) Who can raise questions about the management of the filters and flow of relevant and reliable information (i.e. the issue of the management of conditions of insecurity and uncertainty)? We do not answer these questions, which exceed the scope of the present paper. Time is ripe for some conclusions.

21.5 Conclusions

Security has an increasingly central role in our contemporary information societies. Whether understood in terms of *safety* (i.e. the immediate protection of life integrity and life projects) or *security* (i.e. the mediate protection of the implementation of life projects), security is key policy lever, which allows a government to gain greater powers of direction and control with the aim of immunizing society against the risks that seem to threaten its integrity. However, such risks are never totally immunized by the policies of security, since security is neither managed nor granted on one's own. On the contrary, it is delegated to fiduciary forms that institutionalize and often incorporate into technologically standardized and automated process and devices of risk-management. In this sense, security needs trust and, where trust is, there is risk. For this reason, we should deal with all the challenging aspects and questions remarked in the present paper by bearing in mind that the issue of security does not only concern surveillance, privacy, data protection, and human rights, as everybody repeats, but first and foremost our understanding and construction of trust in the fiduciary forms of risk-management. The way in which these forms of risk-management are institutionalized, structured and deployed through a necessary process of delegation affects the democratic, legal and political legitimacy of the "security issue" in our information societies.

As already remarked, this process of delegation is now closely and strongly entwined with the technological dependency of our society, since it is meant to entrust the management of security to increasingly automated and standardized technological processes and devices. This may entail an incorporation, *by design*, of values and norms in the technological processes and devices, which can bring about invisible abuses, discriminatory effects, as well as unintended consequences affecting individual or collective rights. This may also involve a reallocation of power between different political and social actors and a reconfiguration of institutional arrangements. Furthermore, the impersonal, anonymous, automated and standardized—even though seemingly neutral and benign—technological processes and devices do not always pass through the scrutiny of public understanding and debate that might ensure a more robust political and democratic legitimacy as well as a wider degree of social acceptance.

In our contemporary age, there might be a paradox: security is often invoked as *safety* (against imminent and unpredictable dangers and risks), but most often it is practiced and implemented as *security*, remarkably by the recourse to the available technology in terms of automated and standardized processes and devices. This results—or can result—in a sort of contradiction, for which the contingent (risk and danger) presides, at present, in the implementation and legitimation of what is serial, automated, and standardized. Against this backdrop, the modern project of political and legal construction of *individual identity* (individual rights, freedoms, principles and values) is therefore put at risk and] fated to drown in the opacity of the collective fears more or less artificially constructed by the request and promise of a greater security.

References

Barocas, S., S. Hood and M. Ziewitz. 2013, March 29. *Governing algorithms: A provocation piece*. Available at: https://doi.org/10.2139/ssrn.2245322

Bauman, Z., and D. Lyon. 2012. *Liquid surveillance: A conversation*. Cambridge: Polity Press.

Beck, U. 1992. *Risk society. Towards a new modernity*. London: Sage Publications.

Benkler, Y. 2006. *The wealth of networks: How social production transforms markets and freedom*. New Haven: Yale University Press.

Brey, P. 2010. Values in technology and disclosive computer ethics. In *Information and computer ethics*, ed. L. Floridi. Cambridge: Cambridge University Press.

Busch, L. 2011. *Standards. Recipes for reality*. Cambridge, MA: MIT Press.

Castells, M. 2009. *Communication power*. Oxford: Oxford University Press.

Durante, M. 2007. *Il futuro del web. Etica, diritto, decentramento. Dalla sussidiarietà digitale all'economia dell'informazione in rete*. Torino: Giappichelli Editore.

———. 2010. What is the model of trust for multi-agent systems? Whether or not E-trust applies to autonomous agents. *Knowledge, Technology, and Policy* 23: 347–366.

———. 2012. E-democracy as the frame of networked public discourse. Information, consensus and complexity. In *Legitimacy 2.0. E-democracy and public opinion in the digital age, paper series – 25th IVR World Congress: Law, science and technology*, ed. P. Mindus, A. Greppi, and M. Cuono, 1–28. Frankfurt am Main: Goethe University Press.

Floridi, L. 2010. *Information. A very short introduction*. Oxford: Oxford University Press.

———. 2013. *The Ethics of Information*. Oxford: Oxford University Press.

———. 2014. *The fourth revolution. How the infosphere is reshaping human reality*. Oxford: Oxford University Press.

———., ed. 2015. *The onlife manifesto. Being human in a hyperconnected era*. Dordrecht: Springer.

Giddens, A. 1990. *The consequences of modernity*. Stanford: Stanford University Press.

Hobbes, T. 1991. *Leviathan [1651]*. Cambridge: Cambridge University Press.

Kallinikos, J. 2006. *The consequences of information. Institutional implications of technological change*. Cheltenham: Edward Elgar.

———. 2011. *Governing through technology. information artefacts and social practice*. Houndmills Basingstoke: Palgrave Macmillan.

Lessig, L. 1999. *Code and other laws of cyberspace*. New York: Basic Books.

———. 2006. *Code: Version 2.0*. New York: Basic Books.

Locke, J. 1998. *Two treatises of government [1690]*. Cambridge: Cambridge University Press.

Luhmann, N. 1979. Trust: A mechanism for the reduction of social complexity. In *Trust and power: two works*, ed. N. Luhmann, 1–103. New York: Wiley.

Lyon, D. 2009. *Identifying citizens: ID cards as surveillance*. Cambridge: Polity Press.

Moor, J. 1985. What is computer ethics? *Metaphilosophy* 16 (4): 266–275.

Nissenbaum, H. 2004. Will security enhance trust online, or supplant it? In *Trust and distrust in organizations: Dilemmas and approaches*, ed. R.M. Kramer and K.S. Cook, 155–188. New York: Sage.

Pagallo, U. 2013. *The laws of robots: Crimes, contracts, and torts*. Dordrecht: Springer.

———. 2014. *Il diritto nell'età dell'informazione. Il riposizionamento tecnologico degli ordinamenti giuridici tra complessità sociale, lotta per il potere e tutela dei diritti*. Giappichelli Editore, Torino.

Pagallo, U., and M. Durante. 2016. The pros and cons of legal automation and its governance. *European Journal of Risk Regulation* 7 (2): 323–334.

Zittrain, J. 2010. Perfect enforcement on tomorrow's Internet. In *Regulating technologies: Legal futures, regulatory frames and technological fixes*, ed. R. Bronsword and K. Yeung, 125–156. London: Hart Publisher.

Chapter 22
The Challenges of Digital Democracy, and How to Tackle Them in the Information Era

Ugo Pagallo

Abstract Scholars examine legal hard cases either in the name of justice, or in accordance with the principle of tolerance. In the case of justice, scholars aim to determine the purposes that all the norms of the system are envisaged to fulfil. In the second case, tolerance is conceived as the right kind of foundational principle for the design of the right kinds of norms in the information era, because such norms have to operate across a number of different cultures, societies and states vis-à-vis an increasing set of issues that concern the whole infrastructure and environment of current information and communication technology-driven societies. Yet the information revolution is triggering an increasing set of legal cases that spark general disagreement among scholars: Matters of accessibility and legal certainty, equality and fair power, protection and dispute resolution, procedures and compliance, are examples that stress what is new under the legal sun of the information era. As a result, justice needs tolerance in order to attain the reasonable compromises that at times have to be found in the legal domain. Yet, tolerance needs justice in order to set its own limits and determine whether a compromise should be deemed as reasonable.

Keywords Justice · Tolerance · Hard legal case analysis · Information and communication technologies · Information ethics · Paradoxes of tolerant rules

22.1 Introduction

Today's information revolution should be considered as a set of constraints and possibilities that transform or reshape the environment of people's interaction and their democratic institutions. Whereas, over the past centuries, human societies have been related to information and communication technologies (ICTs), but mainly

U. Pagallo (✉)
Dipartimento di Giurisprudenza, University of Turin, Turin, Italy
e-mail: ugo.pagallo@unito.it

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_22

dependent on technologies that revolve around energy and basic resources, current societies are increasingly dependent on ICTs and furthermore, on information and data as a vital resource. This dependency triggers some basic novelties in terms of complexity and legal enforcement, which impact pillars of the law and democratic processes by reshaping the balance between resolution and representation, as well as the right of the individuals to have a say in the decisions affecting them. Matters of accessibility and legal certainty, equality and fair power, protection and dispute resolution, procedures and compliance, are fruitful examples to stress what is new under the legal sun of the information era. As today's debate on internet governance further illustrates, it is far from clear how we should grasp the model that may successfully orient our political strategy in terms of transparency, justice and tolerance, so as to strike the right balance between people's representation and political resolution (Durante 2015; Pagallo 2015a, b).

However, by examining the legal challenges of the information era, we should avoid a misunderstanding. Many current troubles with democratic processes are often discussed and presented as if they were new, although this is in fact not the case. Think of Milton Mueller's analysis on *Networks and States*, in which one of the main theses is that most discussions of internet governance insist on "the issues of who should be 'sovereign' – the people interacting via the internet or the territorial states" (Mueller 2010: 268). Likewise, contemplate Nafeez Ahmed's account on the "secret network" behind mass surveillance, endless war, and Skynet, so that a secret Pentagon-sponsored group has been using digital technology over the past decades, as a way "to legitimize the power of the few over the rest of us" (Ahmed 2015). Also, reflect on current debate on the lack of transparency and of public consultation that affects both institutions, e.g. the EU Commission, and the transnational governance network that includes such organizations as the International Criminal Court, the International Organization for Standardization (ISO), the World Trade Organization (WTO), and more (Keohane 2003; Castells 2005; etc.). These open issues of democracy can be traced back to the work of the most distinguished Italian philosopher of the second mid twentieth century, Norberto Bobbio. In *The Future of Democracy* from 1984, Bobbio explored what he dubbed the "six broken promises of democracy," which cast light on such crucial aspects of today's discussions that revolve around the respect for individual sovereignty, the primacy of political representation over the protection of particular interests, the defeat of oligarchies, the increase of spaces for self-government, the education of citizens, or the transparency of governments (Bobbio 2014). From this latter point of view, it follows that many problems of current digital democratic trends are as old as democratic theory. How, then, can we distinguish between endurances and discontinuities? And moreover, how should we tackle them?

In order to address this complex set of issues, let us restrict the focus of the analysis on how jurists commonly assess cases of legal disagreement that may potentially concern either the broken promises of democracy or the new challenges of the information revolution. By leaving aside the normative theories of democracy and its justification, the paper does not take into account discussions between instrumentalism and non-instrumental values, the role of democratic citizenship,

multiple versions of democratic authority, or legislative representation. Rather, the attention is drawn to that which jurists usually sum up as their legal "hard cases" (Hart 1961; Dworkin 1985; Shapiro 2007; Pagallo and Durante 2016a). General disagreement may regard the meaning of the terms framing the question, the ways such terms are related to each other in legal reasoning, or the role of the principles that are at stake in the case. Examples of this divergence concern today's clauses of due process, the protection of fundamental rights vis-à-vis matters of national security, mechanisms of legal automation, and so on. These cases are particularly relevant for they trigger a further form of meta-disagreement on how we should grasp the hard cases of the law and hence, how the troubles with digital democracy should be tackled in the information era.

All in all, scholars may examine the legal hard cases either in the name of justice, or in accordance with the principle of tolerance. Let us call them followers of Rousseau and Locke, respectively. In the first case, justice represents the moral principle with which scholars aim to determine the purposes that all the rules of the system are envisaged to fulfil. In the case of tolerance, it is the latter that provides the foundational principle of a fair, peaceful, and democratic society. Each approach has its merits and limits: as to the merits, both stress what current cases of legal and political disagreement may have in common, e.g. the quest for consent as a matter to be evaluated in terms of justice, or of tolerance. As to the limits of each approach, what ultimately is at stake has either to do with the threat of an intolerant justice, or the risk of a toothless tolerance. In order to understand why this may be the case today, let us proceed with the thesis that (also digital) democracy rests on justice and what this means in the information era.

22.2 On Justice and Its Limits

The first way to address the broken promises of democracy and the new challenges of the information revolution regards a popular stance in the tradition of modern political thought: Justice is the moral principle with which scholars aim to determine the purposes that all the norms of the system are envisaged to fulfil. Three centuries after Rousseau's social covenant, and almost two after Kant's, consider a classic text like Rawls' *A Theory of Justice* and, in the legal domain, the idea that a "right answer" can be found for every case under scrutiny. On the one hand, the thesis is that "justice is the first virtue of social institutions, as truth is of systems of thought" (Rawls 1999: 3). On the other hand, Dworkin and his followers have suggested the uniquely right answer-approach. According to this stance, a morally coherent narrative should grasp the law in such a way that, given the nature of the legal question and the story and background of the issue, scholars can attain the answer that best justifies or achieves the integrity of the law (Dworkin 1985). By identifying the principles of the system that fit with the established law, jurists could apply such principles in a way that presents the case in the best possible light.

As an instance of this Dworkinian approach, reflect on some challenges of today's democracy on the basis of a morally coherent theory, such as the ethics of information (Floridi 2013). This level of abstraction represents all the entities and agents in the system, as well as the whole environment, in terms of (not only, but also) meaningful data. Contemplate on this basis the set of problems that regard the legal regulation of extraterritorial conduct in cyberspace, so that, pursuant to the traditional tenets of the rule of law (Bingham 2010), what "the laws of the land" should be often is hard to tell in the new context. Furthermore, even if we may agree on such laws of the land for digital democracy, there is an increasing number of cases in which the law lays down different set of obligations for online and offline interaction. A significant example is given by the right to control communication to the public in the field of copyright law, which "imposes more stringent obligations on the users of cyberspace technologies" (Reed 2012: 194). This creates the potential for litigation over whether "the laws of the land" should apply equally between the real world and another dimension of social interaction, notably cyberspace. Going back to the tenets of information ethics, the overall idea is thus to grasp these legal issues within the normative framework that governs the entire life cycle of information and determines what is right in the "info-sphere." The differentiation between online and offline interaction can be evaluated in a Dworkinian manner, by drawing the attention to the moral laws of information ethics and whether such differentiation prevents either "entropy," i.e. the destruction and corruption of informational objects, or contributes to their flourishing in the info-sphere. The more we deal with ICT-driven societies, the more their legal and political issues become a matter of access to, and control and protection over, information, the more we should pay attention to how to enrich the info-sphere, or prevent cases of informational entropy. Therefore, can a morally coherent theory attain the Dworkinian right answer for all of the ways in which traditional democratic problems have realigned in the information era?

The set of multiple issues that may spark legal disagreement shows a further set of cases in which different moral and political assumptions provide many right answers out there. No algorithm can mechanically be applied to rights and interests that should be balanced in the name of, say, Article 27 of the Universal Declaration of Human Rights (UDHR) on digital copyright and intellectual property, Article 6 of the European Convention on Human Rights (ECHR) on the due process in the information era, or the protection of further fundamental rights, e.g. privacy, vis-à-vis national security and the new frontiers of cyber war (Pagallo 2015c). Even *Law's Empire* seems to suggest this conclusion: "For every route that Hercules took from that general conception to a particular verdict, another lawyer or judge who began in the same conception would find a different route and end in a different place" (Dworkin 1986: 412). By taking into account current debates on internet governance and digital copyright, national security and data protection, and more, it seems fair to admit that no theory of justice can offer the one-size-fits-all answer for the complex set of issues the law faces today. Rather, what these cases illustrate is a class of legal issues that confront us with something new, which requires a reasonable compromise between many conflicting interests. Although this is of

course the stance Herbert Hart made popular with his work, it does not follow that we have to buy any of his theoretical assumptions on, say, the rule of recognition and the minimum content of natural law, to admit that a reasonable compromise has at times to be found in the legal domain (Hart 1961). As previous international agreements have regulated technological advancements over the past decades in such fields as chemical, biological and nuclear weapons, or the field of computer crimes since the early 2000s, many claim that a new agreement on, for example, today's laws of the war, e.g. robot soldiers, is necessary (Pagallo 2013).

The second fundamental moral principle, or Rawlsian virtue of social institutions, seems thus to be tolerance. The latter should in fact complement justice, because an open attitude to people whose opinions may differ from one's own, is that on which any reasonable compromise ultimately relies. Regardless of the field under scrutiny, such as military robotics, data protection, digital copyright and intellectual property, international cooperation, financial regulation, internet governance, and more, let us now explore how far this idea of tolerance goes in the next section.

22.3 On Tolerance and Its Limits

The "tolerant approach" to the current issues of digital democracy can reasonably be traced back to the liberal variants of contractualism, such as Locke's *A Letter concerning Toleration* from 1689, or John Stuart Mill's *On Liberty* (1859). Tolerance represents both a fundamental moral principle of normative design and a key ingredient for such legal hard cases that require a reasonable compromise between many conflicting interests. Tolerance, rather than justice, may provide the right kind of foundational principle for the design of the right kinds of norms in the information era, because such rules have to operate across cultures, societies and states vis-à-vis an increasing number of issues that concern the whole infrastructure and environment of current ICT-driven societies. The more such issues appear "hard," i.e. a source of general disagreement, the more a reasonable compromise should be attained, the more tolerance provides the foundational principle of a fair, peaceful and democratic society (Floridi 2014).

However, it is far from clear how to determine whether or not the compromises that have at times to be found in the legal domain are tolerantly "reasonable." In addition, the open issues of digital democracy raise the old dilemma of how to avoid, or solve, the paradox of tolerance, namely the idea that "unlimited tolerance must lead to the disappearance of tolerance" (Popper 2013). Scholars that insist on the need of some reasonable compromise, have the burden to prove how tolerance can set its own limits without justice. After all, contrary to the latter, which can reinforce itself through its own application, tolerance runs into the problem of its excessive scope. As Popper used to remark time and again, "if we extend unlimited tolerance even to those who are intolerant, if we are not prepared to defend a tolerant society against the onslaught of the intolerant, then the tolerant will be

destroyed, and tolerance with them" (*op. cit.*, 581). In light of current trends on global surveillance, emergency powers and the wave of terrorist attacks that have recommended an intensification of security programs at national and international levels, is there any room for tolerance and its reasonable compromises today? Don't these trends suggest that plans for the transparency of governments, i.e. Bobbio's final broken promise of democracy, will be postponed for quite a long time?

A feasible way out has been proposed by Floridi (2014). Contrary to the traditional idea of tolerance as a dual interaction between an "A" and a "B", he suggests that we should grasp the principle of tolerance as a ternary relation. "A" should not tolerate any "B's ϕ–ing" when "C"–which is significantly affected by "B's ϕ–ing"–does not provide uncoerced and informed consent. According to the traditional point of view, if someone ("A") does not tolerate something ("B's ϕ–ing"), intolerance can be justified because that 'something' (B's ϕ–ing") is deemed as unjust. By grasping the idea as a ternary relation, Floridi claims, "we now have a way of constraining toleration by means of tolerance, without a circular recourse to the principle of justice. The need for interpretation through public debate assumes that, by default, toleration is legitimate and should be exercised whenever it is not constrained by tolerance or unless the interpretation of the conditional convincingly shows otherwise" (Floridi 2014: 23).

Some troubles with the scheme are admitted by Floridi as to, say, the meaning of C to be significantly affected by B's ϕ–ing, or the notion of C's consent. For instance, consider that consent is still a fundamental principle of the EU data protection legal framework and yet, a number of reasons suggest why the notice and consent-approach is under strain: privacy notices are more often labyrinthine and it is hard for individuals to determine long-term risks of their consent, so as to balance them against short-term gains. The 2016 EU new regulation on data protection, the so-called GDPR, significantly puts forward further approaches, e.g. data protection impact assessments and the principle of accountability, in order to properly tackle the challenges of the information era (Pagallo 2017a). But, going back to Floridi's "tolerant approach," how about all the cases in which "C" is a group, or a collective, that is divided about their reaction, or tolerance, concerning "B's ϕ–ing"? What should "A" do? Since "A" has not to take sides in the name of justice, what should A's criteria of tolerance be? Does a single dissident of C preclude A's toleration, or should it be a significant minority? In more general terms, is there a way to avert the conclusion that at times, the tolerant needs to resort to some idea of justice?

The troubles with democracy in the information era have apparently led to a vicious circle. On the one hand, no theory of justice can offer an algorithm to be mechanically applied to all the hard cases of the law and no surprise then, that some present tolerance as the only way to cope with the reasonable compromises that at times should be found in the legal and political fields. After all, the information revolution has produced, and will increasingly raise, cases of general disagreement that concern multiple legal regulations aiming to govern cross-border interaction in a globalized world (Pagallo 2017b). Whilst, since the mid 1990s, states have begun to react to the challenges of the information revolution with the same tools of technology, e.g. by embedding normative constraints into ICTs, this reaction has

triggered additional hard cases and the need for further crucial compromises on, e.g., legal automation. Whether, and to what extent, should the normative side of the law be transferred from the traditional "ought to" of legal systems to automatic techniques through the mechanisms of design, codes, and architectures? (Pagallo 2012; Pagallo and Durante, 2016b).

On the other hand, tolerance has some limits of its own whenever, in Floridi's phrasing, those affected by any "x's ϕ–ing" disagree on whether or not they should provide their consent. Remarkably, this is a key point of Bobbio's broken promises of democracy that some of the new challenges of the information era have brought about as a matter of certainty, equality, and compliance. The less legal boundaries are clear in digital environments, the more this situation may lead to the illegitimate condition where states claim to regulate extraterritorial conduct by imposing norms on individuals, who have no say in the decisions affecting them. This scenario brings us back to (a variant of) our dilemma, i.e. either a toothless tolerance or an undemocratic justice. As a result, is there any feasible way out for this vicious circle between tolerance and justice?

The short answer is "yes." Let me argument why in the conclusions of this paper.

22.4 Conclusions

The aim of this paper has been to examine the ways in which jurists commonly address cases of legal disagreement that may potentially concern both the broken promises of democracy and the new challenges of the information revolution. The first perspective has to do with the popular stance of the modern political tradition, according to which a morally coherent theory could determine the purposes that all the norms of the system are envisaged to fulfil. How this works has been illustrated with the tenets of a morally coherent theory, such as Floridi's ethics of information. By conceiving all the agents and processes of the system in terms of information, the first moral law of this perspective claims that every form of informational entropy, i.e. any kind of impoverishment of being in the info-sphere, ought not to be caused. Moreover, the informational entropy ought also to be prevented or removed. This sort of Dworkinian approach to the challenges of the digital era can be helpful at times. In addition to the principle of equality and an increasing number of cases in which the law imposes different obligations for online and offline interaction–as mentioned above in Sect. 22.2–consider problems of transparency and the protection of privacy and personal data. The tenets of information ethics may provide that sort of moral coherent theory with which to attain a uniquely right answer, e.g. a fair balance between principles and norms that on the one hand constrain the flow of information and on the other, flesh out the factors on which the availability of information, or the conditions of its accessibility, namely individual, social, and political transparency, depend. The focus should be on whether informational entropy is either prevented, or removed, or whether the flourishing of the entities, which are stake with such a balancing, is promoted.

However, pace Dworkin, even Floridi would admit that the moral laws of information ethics cannot provide the uniquely right answer for every legal hard case at hand. This is why, after his informational theory of justice, Floridi has proposed to complement it with the principle of toleration (Floridi 2014). In legal terms, this open attitude to people whose opinions may differ from one's own, has been illustrated with cases of general disagreement on how we should regulate digital copyright, cyber war, national security and data protection, and more. As both a fundamental moral principle of normative design and a key ingredient for how to tackle the hard cases of the law, tolerance paves the way to the reasonable compromises that at times have to be found between many conflicting interests. Yet, the previous section ended with the example of a group, or a collective, affected by a certain "x's ϕ–ing," that disagree on how to react, i.e. whether or not they should provide their consent. If tolerance may need justice, we should avert an intolerant justice and moreover, mere injustice. Therefore, how can we determine whether a certain compromise is reasonable?

After the traditional dual approach to the principle of tolerance and Floridi's ternary relation, let us assume here a third approach. In accordance with another of its meanings, tolerance can be understood as the permitted variation in some measurement or other characteristic of an object or informational entity. On this basis, going back to the moral laws of information ethics and its idea of justice, old and new challenges of digital democracy suggest that we should tackle justice with a margin of tolerance. Although it is in the name of justice that scholars interpret the purposes that all the norms of the system are envisaged to fulfil, justice still needs tolerance, in order to cope with cases of general disagreement that constitute the legal hard cases and its reasonable compromises. So, the more legal and political interaction increasingly revolves around how to monitor, regulate, or control the flow of information in today's ICT-driven societies, the more we should pay attention to the permitted variation in the amount of informational entropy that every reasonable compromise should minimize. The more the informational entropy is reduced or prevented, the more an agreement should be deemed as reasonable. This is the yardstick with which we can both evaluate the hard cases of today's digital democracy, and build a tolerant justice.

References

Ahmed, Nafeez. 2015. *How the CIA made Google* (Part I) *& Why Google made the NSA* (Part II), at https://medium.com/@NafeezAhmed/how-the-cia-made-google-e836451a959e and https://medium.com/@NafeezAhmed/why-google-made-the-nsa-2a80584c9c1. Last accessed 15 Mar 2015.

Bingham, Tom. 2010. *The rule of law*. London: Penguin.

Bobbio, Norberto. 2014. *The future of democracy*. Minnesota: University of Minnesota Press.

Castells, Manuel. 2005. Global governance and global politics. *Political Science and Politics* 38: 9–16.

Durante, Massimo. 2015. The democratic governance of information societies. A critique to the theory of stakeholders. *Philosophy and Technology* 28 (1): 11–32.
Dworkin, Ronald. 1985. *A matter of principle*. Oxford: Oxford University Press.
———. 1986. *Law's empire*. Cambridge, MA: Harvard University Press.
Floridi, Luciano. 2013. *The ethics of information*. Oxford: Oxford University Press.
———. 2014. Toleration and the design of norms. *Science and Engineering Ethics*, (October): 1–29.
Hart, Herbert L.A. 1961. *The concept of law*. Oxford: Clarendon.
Keohane, Robert O. 2003. Global governance and democratic accountability. In *Global governance and democratic accountability*, ed. D. Held and M. Koening-Archibugi. Cambridge: Polity Press.
Mueller, Milton L. 2010. *Networks and states: The global politics of internet governance*. Cambridge: MIT Press.
Pagallo, Ugo. 2012. Cracking down on autonomy: Three challenges to design in IT law. *Ethics and Information Technology* 14 (4): 319–328.
———. 2013. *The laws of robots: Crimes, contracts, and torts*. Dordrecht: Springer.
———. 2015a. The realignment of the sources of the law and their meaning in an information society. *Philosophy & Technology* 28 (1): 57–73.
———. 2015b. Good onlife governance: On law, spontaneous orders, and design. In *The onlife manifesto: Being human in a hyperconnected era*, ed. L. Floridi, 161–177. Dordrecht: Springer.
———. 2015c. Cyber force and the role of sovereign states in informational warfare. *Philosophy & Technology* 28 (3): 407–425.
———. 2017a. The legal challenges of big data: Putting secondary rules first in the field of EU data protection. *European Data Protection Law Review* 3 (1): 34–46.
———. 2017b. The broken promises of democracy in the information era. In *Digital democracy in a globalized world*, ed. C. Prints, C. Cuijpers, P.L. Lindseth, and M. Rosina, 77–99. Cheltenham: Elgar.
Pagallo, Ugo, and Massimo Durante. 2016a. The philosophy of law in an information society. In *The Routledge handbook of philosophy of information*, ed. L. Floridi, 396–407. Oxon/New York: Taylor and Francis.
———. 2016b. The pros and cons of legal automation, and its governance. *European Journal of Risk Regulation* 7 (2): 323–334.
Popper, Karl R. 2013 *The open society and its enemies*. New introduction by Alan Ryan, essay by E.H. Gombrich, single volume ed. Princeton: Princeton University Press.
Rawls, John. 1999. *A theory of justice*. Cambridge, MA: Belknap Press of Harvard University Press.
Reed, Chris. 2012. *Making laws for cyberspace*. Oxford: Oxford University Press.
Shapiro, Scott J. 2007. *The 'Hart-Dworkin' debate: A short guide for the perplexed*, Public law and legal theory working paper series, 77. Michigan Law School.

Correction to: Modal Ω-Logic: Automata, Neo-Logicism, and Set-Theoretic Realism

David Elohim

Correction to:
Chapter 4 in: D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_4

The author's name in Chapter 4 - Modal Ω-Logic: Automata, Neo-Logicism, and Set-Theoretic Realism has been corrected from 'Hasen Khudairi' to 'David Elohim'. The name has been corrected in pages 5, 6, 74, 79 and in Index as well.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-030-01800-9_4

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9_23

Index

D. Berkich, M. V. d'Alfonso (eds.), *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence*, Philosophical Studies Series 134,
https://doi.org/10.1007/978-3-030-01800-9

T