Journal of Artificial Intelligence and Consciousness

The AI Wars, 1950 - 2000, and their consequences

Eric Dietrich

Philosophy Department, Binghamton University,
Binghamton, New York, 13902, USA
dietrich@binghamton.edu

Chris Fields

23 Rue des Lavandières,
Caunes Minervois, 11160, FRANCE
fieldsres@gmail.com

John P. Sullins

Philosophy Department, Sonoma State University,
Rohnert Park, California, 94928, USA
john.sullins@sonoma.edu

Bram Van Heuveln

Cognitive Science Department, Rensselaer Polytechnic Institute,
Troy, New York, 12180, USA
heuveb@rpi.edu

Robin Zebrowski

Cognitive Science Department, Beloit College,
Beloit, Wisconsin, 53511, USA
zebrowsr@beloit.edu

Received Day Month Year
Revised Day Month Year

Philosophy and AI have had a difficult relationship from the beginning. The "classic" period from 1950 to 2000 saw four major conflicts, first about the logical coherence of AI as an endeavor, and then about architecture, semantics, and the Frame Problem. Since 2000, these early debates have been largely replaced by arguments about consciousness and ethics, arguments that now involve neuroscientists, lawyers, and economists as well as AI scientists and philosophers. We trace these developments, and speculate about the future.

Keywords: Architecture; consciousness; ethics; Frame Problem; neuroscience; phrōnesis; semantics; Turing test.

1. Introduction

Between the publication of Turing's "Computing machinery and intelligence" in 1950 and the gradual thawing of the deep AI winter of the 1990s, four distinct philosophical arguments targeting the very foundations of AI as an enterprise arose, briefly commanded enormous attention, then subsided. The first of these "AI Wars" began with Lucas [1961], and questioned the logical cogency of AI. The second can be dated from Minsky & Papert [1969] and involved arguments between partisans of different architectures. The third was launched by Dreyfus [1972] but became most intense following Searle [1980]; it attacked the possibility of AI systems having semantics. The fourth, and in many ways most subtle, began with McCarthy & Hayes [1969] and concerned the meaning and implications of the Frame Problem. The AI wars of this period influenced, and were influenced by, the in-hindsight ludicrous over-optimism of the early 1960s, the Lighthill [1973] Report and other disasters of the 1970s, the enthusiasm around expert systems of the 1980s, and the gradual transitions toward artificial neural networks (ANNs) and applied robotics in the 1990s. By the early 2000s, the philosophical conflicts of AI's first 50 years were largely over. They ended not in victory for either side but in stalemate. In their place, new debates have arisen, about the nature of consciousness, and about both the ethics of AI and the possibility of AI systems themselves being ethical. Turing [1950] foresaw both of these developments, though he discounted their relevance to "intelligence" as he construed it. Here we ask why this happened, and how it affected both AI and philosophy.

Touching as they did on fundamental issues of metaphysics, epistemology, and the philosophy of mind, language, and science, the AI Wars appeared to those involved to be restructuring philosophy itself. Sloman [1978] confidently predicted, for example, that "within a few years, if there remain any philosophers who are not familiar with some of the main developments in artificial intelligence, it will be fair to accuse them of professional incompetence" and that considerations of AI systems would pervade the teaching of academic philosophy. This clearly did not happen: many if not most philosophers remain ignorant of AI, and most philosophical pedagogy never mentions it. Our new book, *Great Philosophical Objections to Artificial Intelligence: The History and Legacy of the AI Wars* (Bloomsbury, 2021), is an attempt to remedy this situation. We summarize its main arguments here, while adding some deeper analysis more accessible to fellow researchers.

The debates about AI consciousness and ethics that have arisen since 2000 involve neuroscientists, lawyers, and economists, among others, as well as AI researchers and philosophers. As were the "classic" AI Wars of 1950–2000, these new conflicts are part of the general cultural conflict that began with the Scientific Revolution. Their urgency derives in part from the growing realization that the post-industrial economy is rapidly becoming the AI economy, and in part from a concern that the post Cold War geopolitical system may become an AI-driven political system. These debates challenge an assumption that seemed natural in the

previous era of AI: the assumption that AI's goal – or in more current language, AGI's goal – is a machine with *human-like* intelligence. They suggest that this previous goal is obsolete, and that a different and much more radical goal waits on the verge of formulation. They point toward an era of opportunities, uncertainty, and far higher stakes than the initial, 20^{th}-century, era of AI.

2. The First War: Is AI Even Possible?

2.1. *Lucas' argument from Gödel's theorem*

The first serious argument against the logical possibility of AI, Lucas [1961], was presented to the Oxford Philosophical Society in 1959, nine years after Turing [1950] and three years after the informal debut of Logic Theorist [Newell, Shaw & Simon, 1957] and John McCarthy's introduction of the term "artificial intelligence" at the 1956 Dartmouth Workshop on AI. Ironically, the formal publication of Lucas [1961] coincided with that of the General Problem Solver [Newell & Simon, 1961], the first attempt to achieve artificial general intelligence.

Lucas' argument is, in brief, that the logical possibility of AI is inconsistent with Gödel's [1931] incompleteness theorem. Gödel's theorem requires any formal system with at least the power of arthmetic to contain true but unprovable sentences. Lucas argues that a human (in Lucas [1961] a "man") can *recognize* true but unprovable sentences, while a (deterministic) machine can recognize "truth" only through the surrogate of provability in whatever formal system the machine implements. Hence machine intelligence is by Gödel's theorem incomplete, while human intelligence is not.

Lucas [1961] evinced a storm of criticism from philosophers, logicians, and some computer scientists (see Lucas [1996] for a lengthy bibliography). We focus in Dietrich *et al.* [2021] on just one of the many critiques: that Lucas [1961] effectively ignores the distinction between object- and meta-languages, and begs the question against AI by assuming that humans, but not machines, can change the level of language that they are employing in order to "see" that an unprovable sentence is true (or false). Indeed Lucas begs the question explicitly in his final sentence: "There is no arbitrary bound to scientific enquiry: but no scientific enquiry can ever exhaust the infinite variety of the human mind" [Lucas, 1961, p. 127]. If the human mind indeed has "infinite variety" human-like AI is obviously impossible. That no actual argument has been given for "infinite variety" was already noted by Turing [1950] in criticizing an anonymous version of Lucas' argument (p. 445).

Lucas [1996], a "retrospective" on Lucas [1961], discusses the issue of "infinite variety" at greater depth. Here Lucas acknowledges the possibility of shifting language levels, but insists that while an AI system might do this any finite number of times, a human has "infinite potentiality" for recognizing unprovable truths. Indeed "infinite potentiality is an essential part of the concept of mind" [Lucas, 1996, p. 109], a statement that goes some way toward explaining why neither AI researchers nor cognitive psychologists have, by and large, taken Lucas' arguments seriously.

But while Lucas admits that this argument from an intuition of unlimited cognitive power is "hand-waving" [Lucas, 1996, p. 105], he clearly means it seriously.

What is going on here? Lucas clearly believes that humans are special, that they are unbounded in some way. This sense of unboundedness reappears in the arguments of Dreyfus [1972] and Searle [1980] that humans have an access to meaning that AI systems can never, even in principle, achieve. It appears again in the anti-AI arguments of Penrose [1989], who dresses Lucas' argument from Gödel's theorem in the language of quantum theory (see also Penrose [1994] for further elaboration). All of these arguments are intuition pumps, stoking the feelings that we *know* how our minds work and that we *know* that they do not work like machines. How do we know these things? Because we are conscious. Lucas' argument is, we argue, at bottom not an argument about cognition, but an argument about consciousness. Humans have "infinite potential" because they are conscious; machines have only finite potential because they are not. We will return to this question of consciousness in §4 and §6 below. The beginning of the argument about AI and consciousness was not, however, Lucas [1961] but rather Turing [1950].

2.2. *The Turing test*

Turing [1950] asks "can machines think?" (p. 433). Turing dismissed arguments that machines could not think because they could never be conscious, pointing out correctly that we only *assume* that other humans are conscious. He famously offered his "imitation game" or "Turing test" as a purely behavioral test, applicable to any system, human or machine. The Turing test is widely regarded, especially in popular media, as a criterial test for machine intelligence: if the interrogator cannot tell the difference between the machine's responses and those of a human, then the machine should be declared intelligent. Some even see the Turing test as some kind of operational definition of intelligence. The Loebner competition operationalized the test with prizes, and has even declared that comically-incompetant chatbots have "passed" the test (see Aaronson [2014a] for an example).

We argue that Turing did not intend the imitation game as a serious test for or definition of intelligence, and that to read Turing [1950] as if he did so is a mistake. First, as a measure of intelligence it does not pass any kind of scientific muster. It is sloppy and subjective (Who is the interrogator? What is the conversation about? How long is the conversation?), needlessly indirect (Why not interact directly with the subject in order to gauge its properties?), narrowly focused on human-level intelligence (What about non-human-level intelligence?), and it does not take into account the fact that intelligence is an umbrella term for a multi-faceted concept. While Turing made suggestions about length of interaction (five minutes) and topic (anything), these are unmotivated and seem off-hand. Even in Turing's time there were more objective tests for gauging various aspects of intelligence; Turing offers no discussion of these.

More important, however, is Turing's use of the game as a rhetorical ploy, e.g. in

his discussion of consciousness. At the start of his paper, Turing makes it abundantly clear that he does not believe there is any kind of clear and objective definition or measure of intelligence. Rather, Turing expresses the belief that our concepts and attributions of "intelligence" would gradually change over time, possibly as a result of the very coming of more and more cognitively powerful machines. He also, very clearly, believes that our judgements of intelligence are biased towards systems like us, i.e. humans and maybe other mammals. Turing [1948] describes an "experiment" demonstrating that those who would be exposed to the inner working of a machine would be less likely to attribute intelligence to the machine than those who merely saw the machine's input-output behavior. To Turing, this was a demonstration of "intelligence" being a highly value-laden "emotional concept": that many humans tend to think of intelligence being some "special ingredient" that goes beyond mere "cogs and wheels" (perhaps even amounting to "infinite potentiality"). Turing [1950] seems to have introduced the imitation game as a similar thought experiment to make people think twice before quickly dismissing the idea of machine intelligence on the basis of mechanisms merely performing "unimaginative donkey-work". Indeed, if the Turing test is a "test" at all, it is a test of human thinking about – and possible prejudice against – machine intelligence.

Turing has often been criticized for believing that machines would be able to "pass the test" around the year 2000, which seemingly didn't come to pass. However, Turing correctly foresaw the changes in our attitudes regarding the idea of machine intelligence. If anything, Turing probably underestimated how readily we would ascribe cognitive and mental attributes to the machines around us. This raises its own issues, some of which we discuss in §7.1 below. The space of possible intelligences is vast, and human cognition is only a small part of it. The cognitive abilities of our current technologies likely occupy quite different regions of that space, and we should not assume that machines think like us, or that they have the same interests or goals. We discuss these issues of trust further in §7.2.

3. The Second War: Architectures for Intelligence

3.1. *How computer science saved the mind*

Many view AI as somehow anti-human, and AI researchers as partisans of the mechanistic forces of darkness. Turing [1950], not surprisingly, already saw this; he makes fun of this view as the "Heads in the Sand" objection to machine intelligence. More recent versions are easy to come by, e.g. AI is a "a serious danger to a properly human mode of existence" [Madison, 1991, p. 117].

Such objections misrepresent history. Minds were banished from psychology in the early 20^{th} century by the behaviorist revolt against 19^{th}-century speculative mentalism. They were re-introduced when behaviorism was shown to be insufficient even for the description of behavior [Chomsky, 1959]. Two things happened between the banishing of minds from psychology and their re-introduction: 1) mechanisms were invented that could perform arbitrarily complex symbol processing, and 2)

Church [1936] and Turing [1936] proved that arbitrarily many different mechanisms could perform any given symbol processing operation. Computer science, in other words, provided the concept of a *symbol processing mechanism* that psychology needed to talk about minds, a concept formalized by Newell [1980] as the Physical Symbol System hypothesis. Cognitive psychology as we know it was born.

The idea of symbol processing mechanisms had, however, an even more dramatic effect than the re-introduction of minds to psychology. It allowed *all* mechanisms to be seen as symbol processors. One could now talk about biological processes as computations [Turing, 1952] and genes as a code [Zuckerkandl & Pauling, 1965]. The Church-Turing thesis revealed symbol processing everywhere. It enabled the panpsychist revolution of the early 2000s [Strawson, 2006; Goff, 2019, see §6].

3.2. *Symbols versus subsymbols, boxes versus robots*

From the perspective of the Church-Turing thesis, the war over architecure seems strange: Church-Turing tells us that intelligence is implementation independent. Even our best sciences started out getting things wrong, and how they got things wrong is revealing. The second war was about which architecture for intelligence was the right one. There were four main contenders:

(1) Sentence processors
(2) Connectionist and artificial neural networks (ANNs)
(3) Embodied, situated cognitive systems (i.e. robots)
(4) Dynamical systems

Other, more minor contenders are sometimes mentioned, e.g. artificial life programs. We focus on the above in Dietrich *et al.* [2021].

Symbol processors operating on sentence-like representations dominated AI from Logic Theorist through the expert-systems movement of the 1980s [Schank & Abelson, 1977, provides an introduction]. Fodor [1975] and Newell & Simon [1976] state their manifesto: model cognition as it appears to the conscious cognizer, as an inferential process using, roughly, first-order logic with modal and temporal extensions. Their primary weakness is brittleness [Lenat & Brown, 1983], though as we discuss in §6 below, extracting knowledge from experts also proved far harder than anticipated.

Connectionist systems employing networks of simple processors date back to McCulloch & Pitts [1943], but suffered a hiatus following Minsky & Papert [1969]. Their manifesto calls for "subsymbolic" inference, inference below the level of natural language, and learning instead of programming [Rumelhardt & McClelland, 1986; Smolensky, 1988]. In the form of multi-layer deep learning systems [LeCun, Bengio & Hinton, 2015], they dominate applied AI today. Their primary weakness is explainability [Taylor & Taylor, 2020].

Robots were the first AI systems – actually, fantasies of AI systems – to capture the public imagination; Asimov [1950] provides an early example. Fantasies of

robots powerfully motivated the expectation that AI systems would directly model humans (see §6) and left the public drastically unprepared when real robots began to infiltrate the economy (see §7.1). The robotics manifesto is Brooks [1991]; see Anderson [2003] for a broader treatment of embodied AI. Like ANNs, robotics often calls for "subsymbolic" computation, which is often confused with nonsymbolic computation, an oxymoron. The primary weakness of robotics is complexity. Robotics brings AI face to face with the Frame Problem [McCarthy & Hayes, 1969], arguably the hardest problem yet discovered by AI (see §5).

Dynamical systems researchers, by and large, denied the fundamental premise that thinking is computing [van Gelder, 1998, provides an example], and hence effectively (though presumably unknowingly) denied the Church-Turing thesis. The primary weakness of this approach is that it has never produced a working AI system.

Discounting dynamical systems, the three contending architectures have, effectively, divided up thinking into three incompatible research projects: representing and inferring, learning and categorizing, and sensing and moving while avoiding obstacles. Humans, on the other hand, seem to either have one architecture that can do it all, or several architectures that communicate and work well together. The need for integrated architecures was recognized early [Newell, 1990]. Developers responded with systems such as CLARION [Sun, 2007] and LIDA [Franklin *et al.*, 2014]. The deeper question, however, remains: why was there an architecture war in the first place, and why did so many view choice of architecture as a *philosophical* issue? The answer, we suggest, is that the contending architectures involved differing assumptions about the semantics of mental representations.

4. The Third War: Mental Semantics and Mental Symbols

Beginning with Frege [1892] and continuing through the eras of logical empiricism and ordinary-language analysis, philosophers wrestled with the question: "How does language relate to the world?" This question was broadly interpreted as "How does language get meaning (semantics) from the world?" Fodor [1975] encapsulates the consensus answer of 1970s cognitivism: "Language does not relate to the world directly. Language relates to the mind and mind relates to the world." Roughly, the human mind exchanges information with the world via sensors and effectors; this in turn informs a "language module" that produces language.

The third war undermined this intuitively-appealing consensus. It was a complex affair, involving internal conflicts in philosophy [e.g. Dummett, 1978], psychology [e.g. Gibson, 1979], and even biology [e.g. Maturana & Varela, 1980]. We focus in Dietrich *et al.* [2021] on the third war as it affected AI. The attack was lead by Searle [1980] and Chalmers [1996]. Its result was to inextricably link meaning and consciousness, by arguing via meaning that conscious thinking is not explanatorily reducible to physical processes. Here is a summary of the argument, starting with some definitions:

(1) A *process* is anything that occurs through time.
(2) A process X is *explanatorily reducible* to a set Y of subprocesses if and only if explaining X requires analyzing X into its constituent subprocesses, collected in Y, in such a way that some thinking thing can see how the behavior of the processes in Y must result in X.

Science is full of examples of reductive explanations; photosynthesis and its detailed chemical explanation provides an example. Note the crucial role of a "thinking thing" in the definition. By seeing how the processes in Y explain X, the thinker *understands* how and why X works like it does. Understanding is assumed to involve, and require, consciousness. Not all good, useful explanations are reductive, but reductive explanation is the operational goal in the sciences.

(3) A process is (currently) *inscrutable* if (1) we do not have any reductive explanation for it, and (2) we currently have no idea how for find or construct such an explanation. A process I is *inscrutable from process* P if knowing the details of P leaves I inscrutable.

"Knowing" here is also assumed to involve, and require, consciousness.

Searle's [1980] Chinese Room Argument, one of the most famous attacks on AI, concludes that the contents of thinking (i.e. conscious thinking) are inscrutable from any symbolic processor level. Searle's argument has never been conclusively refuted or explained away, even though mountains of analyses and comments have been written on it, as we review in detail in Dietrich *et al.* [2021]. Its conclusion remains if we replace a symbol-manipulating computer with a neural network or an embodied, situated robot cognizer. The contents of thought are, if Searle is right, inscrutable from the level of the computational details of the network or the robot.

Consider an extension of Searle's argument to the human brain: the Fantastic Voyage Argument outlined in Dietrich *et al.* [2021]. Imagine that you are in charge of a large army of molecular-sized people shepherding all the electrochemical information around in the brain. To do this job you have to know everything about the brain at the neural-chemical level. Would such knowledge allow you to know what the brain is thinking? Clearly not. Conclusion: Thinking is inscrutable from the level of the brains neurons. A weaker version of this argument gives the same result. Imagine only that you and your army are witnesses to all that goes on in the brain. Would you know what the brain is thinking? No. The conclusion of the Fantastic Voyage Argument is that the contents of thinking are inscrutable from the level of the brain. Nothing about a thinker's computational or neural details seems even remotely relevant to contentful thinking or conscious thinking. So the needed understanding remains unrealized and for all we know now, unrealizeable.

Both of these arguments depend on the assumption that thinking, knowing, and understanding require consciousness. This suggests, at least, that consciousness is what makes contentful thought inscrutable: that what is really inscrutable is consciousness itself. This conclusion is exactly in line with that of Chalmer's "zombie"

argument [1996], a dramatization of the traditional skeptic's argument against the assumption that other people have minds. The inscrutablity of consciousness suggests, in turn, that it is not reducible but rather fundamental. The natural and increasingly-popular conclusion is some form of panpsychism [Strawson, 2006; Goff, 2019].

The third war raises the cautionary flag already raised by the Turing test. Relying on intuitions about thinking, knowing, and understanding makes these abilities seem inexplicable. Perhaps, as Turing [1950] suggested, our intuitions are not a good guide to how we work.

5. The Fourth War: Rationality, Relevance, and Creativity

We live in a changing world. Even if we just sit on a rock, doing nothing, we, the rock, and the whole rest of the universe changes – restlessly, ceaselessly. This changing would not be a problem without minds. It is only minds that care about change. Minds must keep track of changes in their environments in order to survive, mate, create, and flourish. The rub is that minds are not only finite, they are small compared to the quantity of changes that must be tracked. This is clearly a problem: there are too many changes to keep track of, but not keeping track is a good way to die. This problem contributes to the Frame Problem [McCarthy & Hayes, 1969], the deepest problem yet discovered by AI. But it is not itself the Frame Problem.

Suppose you boil water for making rice using a stove and a metal pot and lid. When the water boils, the water and the pot are hot. To avoid burning your fingers when you lift the lid off, you must remember to update the temperature of the lid: the lid is now hot, too. As you make your rice, you want to update all and only those facts, such as the temperature of the lid, that need updating. But how do you know which facts need updating? You must update your estimation of the temperature of the lid, but you needn't update the number of moons Earth has. Boiling water on a stove does not affect and is not relevant to the number of moons. Obviously, you don't have time to check out all the things that don't need updating – there are far too many of them. But in among all the things that don't need updating *prima facie*, there are some that most definitely need updating. These are the unanticipated side-effects of your action, in this case, of your boiling water for rice. So, are you doomed to go through your vast list of *prima facie* irrelevant things looking for the unanticipated relevancy and update it? When you make a change, what you want is a way of tracking all the facts relevant to the change while ignoring all the facts irrelevant to the change. The quest to find such a way is the Frame Problem. Hayes [1987] put the problem this way (p. 125):

> One feels that there should be some economical and principled way of succinctly saying what changes an action makes, without having to explicitly list all the things it doesn't change as well; yet there doesn't seem to be another way to do it. That is the frame problem.

Hayes was not, in this passage, optimistic about solving the Frame Problem.

Fodor [1987] explained why such optimism is elusive, pointing out that what we need is a robust, predictive theory of what is relevant to what. Such a theory should be expressed formally, in mathematics and logic. So what we want is a (p. 147, emphasis in original):

> FORMAL EXPRESSION [of] our most favoured [and most strongly supported] inductive estimate of the world's taxonomic structure. Well, when we have [such a theory and such a formal expression], most of science will be finished.

In short, we have to wait until science is finished (whatever that means) to solve the Frame Problem. In anything but a closed, artificial domain, this is clearly going to be a long wait, regardless of what "finished science" means. The Frame Problem looks intractable, at least in open domains.

We now know that the Frame Problem is undecidable in any open domain. It is equivalent to the Halting Problem [Dietrich & Fields, 2020]. We have, therefore, but one option when we change something: examine the "obvious" potential side-effects as best we can, and hope that the hidden, unanticipated side-effects are not disastrous. For small changes, this often works. But in general, no. We humans, the smartest beings on the planet, do not actually solve instances of the Frame Problem. Rather, we examine what we can and hope for the best. We are doomed to miss a side-effect here and there. Hence we have global warming, antibiotic resistance, vaccination denial, evolution denial, smoking related diseases, the Challenger space shuttle disaster, the Columbia space shuttle disaster, the Covid-19 pandemic, and on and on.

The Frame Problem is the most important problem yet discovered by AI because it shows us something universal: the consequences of our necessarily limited human knowledge in a vast, strange universe. It shows us, in short, that we do not have "infinite potential": it is irrational to believe that our limited capacities even assure our survival. With the Frame problem, the classic AI Wars come full circle.

6. Neuroscience and the Consciousness War

The widespread adoption of real-time neuroimaging techniques, particularly fMRI, in the early 2000s profoundly altered the academic debate about consciousness. What had historically been a philosophical conversation conducted with thought experiments became an empirical investigation conducted with real experiments and with the neural (activity) correlates of consciousness (NCCs) as the main results being sought [Rees, Kreiman & Koch, 2002]. This empirical turn effectively replaced ontology with functional correlation, a replacement still subject to vigorous debate (see, e.g. Dehaene, Lau & Kouider [2017] versus Carter *et al.* [2018]). More importantly, the focus of the investigation changed. "Unconscious" was clarified to the operational goal of surgical anaesthesia: no reportable or detectable awareness.

What counted, in practice, as reportable or detectable awareness became a critical theoretical issue with immediate implications in the clinic [e.g. Boly *et al.*, 2013]. This focus on awareness, however minimal, as definitive of consciousness is consistent with a literal reading of the notion of a philosophical zombie discussed in §4. It rejects possibly human-specific nuances, e.g. the rich autonoetic awareness of episodic memories [Suddendorf & Corballis, 2007], as definitive of or criterial for consciousness. It focuses instead on easily-universalized criteria like sensitivity to light or the ability to feel pain.

A second critical change, not uncoupled from the first, also occurred in the early 2000s: the Cartesian presumption against nonhuman consciousness was abandoned seemingly overnight. This shift in attitudes occurred, in part, as well-established anatomical similarities across mammalian brains were linked to functional similarities and then, in many cases, to NCCs. Arguments from psychologists and neuroscientists that nonhuman animals [Panksepp, 2005] and even human infants [Rochat, 2003] were aware, experiencing beings were contemporaneous with bold panpsychist statements from philosophers [Strawson, 2006]. Today it is uncontroversial to talk about birds [Güntürkün & Bugnyar, 2016; Nieder, Wagener & Rinnert, 2020] and cephalopods [Mather, 2019; Schnell *et al.*, 2020] as conscious, and a substantial emerging literature extends consciousness to plants [Gagliano & Grimonprez, 2015; Calvo, Baluška & Trewavas, 2020] and even micro-organisms [Levin, 2019; Lyon, 2020]. It is hard to overstate the importance of this shift in intuitions about consciousness in nonhumans: it renders zombie-style arguments (*cf.* §4) far less effective, not only within the bioscience community but for an increasingly-larger public. Thinking of machines, or at least machines with organism-like feedback-driven control systems (and hence positive integrated information Φ [Oizumi, Albantakis & Tononi, 2014]) as conscious becomes considerably more plausible [Tononi & Koch, 2015]. A broad panpsychism becomes, correspondingly, a more attractive default position [Goff, 2019].

A third, somewhat more subtle, change accompanied these first two. Driven in part by the same intuitions that drove the moves toward embodiment and ANNs discussed in §3.2, cognitive psychology increasingly turned toward a deeper investigation of highly-automated "unconscious" thinking [Bargh & Ferguson, 2000]. The ability to perform, and in some cases substantially complete, complex, cognitively-demanding tasks in a highly-automated fashion correlates broadly with expertise in those tasks [Bargh & Ferguson, 2000; Cianciolo *et al.*, 2006], and the difficulty in accessing the "tacit" or "background" knowledge enabling such performance goes some way toward explaining the failure of the knowledge engineering and expert systems paradigms and the resulting AI winter of the 1990s. Social expertise provides an important case in point [Bargh *et al.*, 2012; Chater, 2018]; for example, the rules governing fluent natural language performance, something every human infant learns, remain unknown in detail despite decades of effort. Reasonably fluent natural language systems are, in practice, developed not by explicit rule encoding but by deep learning [Floridi & Chiriatti, 2020], as are in-fact expert AI systems

in domains like protein folding [Senior *et al.*, 2020]. It is not surprising, from this perspective, that explainable AI (XAI) presents many of the same challenges as cognitive psychology [Taylor & Taylor, 2020].

These three developments pose a fundamental challenge to AI as it was conceived by Turing [1950], as it continued to be conceived during the AI Wars of the latter 20$^{\text{th}}$ century, and as it is still conceived in popular media and the popular imagination. The very structure of Turing's imitation game (§2.2) assumes an artificial agent with human-like thought processes, and "human-like" AI remained the target, explicitly or implicitly, throughout AI's first 50 years. The gradual abandonment of this target since 2000 is reflected in our professional language: the quest for human-like AI has been specialized to a subdiscipline: artificial *general* intelligence (AGI; [Goertzel, 2014]). Most of AI, including virtually all of commercial AI, plows forward with little or no concern for the requirements of human-like AGI.

What does this mean for the debates about representational semantics and pragmatic meaning that constituted the Third War (§4), and what does it mean for the "consciousness war" that continues today? Intentional stances taken for convenience aside [Dennett, 1987], all but the most fervent panpsychists regard thermostats as unconscious (see, e.g. the celebrated debate between Aaronson [2014b] and Tononi [2014]). Why? Does anyone care whether thermostats are conscious? Modeling human-like emotion has become important enough to merit its own IEEE journal (Picard [2010] provides a brief history), while intrinsic motivation, e.g. for learning, is a critical requirement for autonomous robotics [Oudeyer & Kaplan, 2007]. Do we care, however, whether an elder-care robot really *feels* sympathy or an iCub really *feels* curious? It is deeply irrational not to care whether autonomous vehicles have functional situational awareness (see §7.1 below), but does anyone really care whether they can *see*?

These questions sound strange, but also strangely familiar. Do we care whether bacteria are conscious? What about trees, insects, or octopi? Within an architecturally-based theoretical framework like IIT [Oizumi, Albantakis & Tononi, 2014], two systems can be constructed of exactly the same kinds of components, and display exactly the same behaviors, but one (with recurrent architecture) has robust consciousness and the other (with feedforward architecture) has none whatsoever. Is it plausible to really care this much about *architecture*?

Questions about caring pull in two directions, that of moral concern and that of ontology. The Cartesian worldview tied these neatly together: having souls made humans ontologically special, and only humans, because of their specialness, were subjects of moral concern. The revolution in neuroscience and the accompanying extension of consciouness beyond humans to other mammals and then to all of life cut this Cartesian knot. It also raised the standard from plausibility to empirical support. Empirical support is conceptually straightforward, though technically challenging, for NCCs. Despite high-profile claims to test "theories of consciousness" as if they were ontological [Reardon, 2019], however, the argument that ontological

claims about consciousness are empirically empty (*cf.* the discussion in §4) steadily gains support [Doerig *et al.*, 2019; Kleiner & Hoel, 2020]. Whether panpsychism is true or false, or true in some limited domain and false outside of it, remains a philosophical question, one increasingly tied back to practical and moral concerns (e.g. by Goff [2019]). We consider these concerns in the next section.

What remains to AI are two questions rarely formulated before 2000, but increasingly urgent today. First, do we even want AIs with human-style consciousness? Human-style consciousness is plentiful: all humans have it. Do we need more of it, packaged in a different form? If so, what for? Second, what might a *native* AI consciousness look like? An octopus, with its highly-decentralized brain, may experience the world very differently from a human. Might octopus-like consciousness be better than human-like consciousness for, say, an air-traffic control system? What kind of consciousness would be optimal for an internet load-balancer, an automated securities-trading system, a mining prospector to be sent to Mars? These questions duck the issue of ontology. But they may be the important questions going forward.

7. AI Applications and the Ethics War

7.1. *Ethical issues surrounding AI applications*

Concerns about the social impact of AI were largely left to science fiction during the period of the classic AI wars, with stories of AI and robot revolutions becoming a common trope. From a philosophical perspective, AI ethics grew out of the field of computer ethics that emerged in the mid-80s [Johnson, 1985/2001]. AI ethics developed as its own research program around 2000; the field of robot ethics soon followed [Bostrom, 2003; Brooks, 2003; Veruggio & Operto, 2008; Wallach & Allen, 2008; Anderson & Anderson, 2011; Lin, Abney & Bekey, 2011; Lin, Abney & Jenkins, 2017a]. This relatively late attention to AI ethics can be understood, at least in part, as due to the paucity of practical AI applications prior to 2000. An opportunity was missed, however, to address potential ethical issues early on when changes in the technology could have been more easily accomplished rather than waiting for problems that may be deeply rooted in the design or implementation of the technology to emerge [Bunge, 1977; Drozdek, 1992; Adam, 1998; Dennett, 1998; Sullins, 2005; Floridi, 2008; Hall, 2009; Moor, 2011; Lin, Abney & Jenkins, 2017a; Bonnemains, Saurel & Tessier, 2018; Coeckelbergh, 2020].

One of the first philosophical questions asked in AI and robot ethics initially arose in science fiction: what is the moral status of AI applications? We consider this question in §7.2 below; here we focus on the more immediately relevant question of impacts on human populations or society. Four areas in particular have received sustained attention:

(1) *Weapons:* The AI wars went from metaphorical to literal over the last few decades as the technology became used in defense initiatives [Lin, Bekey & Abney, 2008; Singer, 2009; Arquilla, 2010; Dabringer, 2011; Floridi & Taddeo, 2014;

Doubleday, 2019]. Governments and industry have moved from remote control systems, to semiautonomous systems, to fully autonomous weapons systems; all three are found regularly on the battlefield and most military systems can switch between all three modalities as the mission dictates. While this advance in technology can seem predetermined and impossible to roll back, it raises significant moral questions, one of the most important being who is responsible when an autonomous system kills [Dennett, 1998; Champagne & Tonkens, 2015; Docherty, 2015; Gunkel, 2017]. It has also been argued that using robots in warfare is possibly even morally preferable to the warfare of the past, which we should remember was never morally unquestionable [Arkin, 2008; Lin, Bekey & Abney, 2008; Arkin, 2009; Sullins, 2010; Lin, 2014]. None of this is settled at this point and there has been vigorous debate and calls for a ban on developing and deploying these systems [Asaro, 2012; Wallach & Allen, 2013; Arkin, 2015; Howard, 2015; Sharkey & Sharkey, 2019].

(2) *Vehicles:* Autonomous vehicles have been a dream since the early days of automotive technology. Much time and resources have been delegated to this industry and there are many prototypes operating on the streets of the world. This technology has more than theoretical ethical impacts: it can impact bystanders with lethal force and it has done so already [Griggs, 2018]. Determining responsibility when things go wrong is an ethical concern with these technologies, as it is with autonomous weapons [Bonnefon, Shariff & Rahwan, 2016; Lin, 2016; Bhargava & Kim, 2017; De Sio, 2017; Lin, 2017]. Since these vehicles are on public streets they represent a public health concern [Fleetwood, 2017]. For all of these reasons, calls for developing governance policies for autonomous vehicles have been raised [Anderson *et al.*, 2014]. An interesting development is the research being done to implement ethical reasoning in the vehicles themselves so they can solve problems as they arise in real time [Goodall, 2014; Gerdes & Thornton, 2015]. The debate about autonomous decision making has traditionally centered around no-win decisions exemplified by trolley problem [Foot, 1967/2002]. Much has been written about this classic philosophical problem [Goodall, 2016; Lin, 2016; Govindarajulu & Bringsjord, 2017]; however, others have argued that it is a red herring and there are other more pressing concerns in programming ethical reasoning in autonomous vehicles [Gurney, 2015; Himmelreich, 2018].

(3) *Social robotics:* Digital assistants, care robots, robot pets, robot sex toys, and even the ubiquitous telephone chatbots intrude into and hence affect ordinary social interactions. Two broad groups of issues have received particular attention: the tendency of these systems to reflect and hence reinforce racial, gender, class, and other stereotypes [Adam, 1998; Noble, 2018; Benjamin, 2019] and their general affect on moral behavior, particularly in the case of sex robots [Levy, 2009; Sullins, 2012; Danaher & McArthur, 2017]. These issues are obviously linked, and concerns within the industry have become sufficient that tech

workers are seeking to unionize to gain more say in their organizations about what projects to accept and to mitigate ethical impacts that their work creates [Gahffary, 2021]. In the particular case of sex robots, some argue that use of such systems can be healthy or even therapeutic [Hawkes & Lacey, 2019], while others argue that they are perverse extensions of sexual violence and abuse that women have suffered for millennia [Richardson, 2016; Kubes, 2019]. A third position has also developed arguing that these applications pose potential moral harm, but if different, more inclusive and ethically bound design parameters are used, they might result in positive additions to the human moral landscape [Sullins, 2012; Peeters, 2019].

(4) *Privacy and governance:* Issues of personal privacy, intrusive marketing, AI-driven manipulation, and surveillance have lead to regulations in some jurisdictions (e.g. the General Data Protection Regulation in the E.U.) and calls for broader legal governance [Etzioni & Etzioni, 2017; Dignum, 2019]. Some argue that laws and governance will be sufficient to maintain control over AI systems and the companies that produce and deploy them [Kowert, 2017], while others point to the widespread use of surveillance for social control [Leong, 2019; Bartoletti, 2020]. More subtle issues include predictive policing [Benjamin, 2019] and the ubiquitous use of "nudging" to construct social-media echo-chambers [Jodi, 2015; Roth, Maziéres & Menezes, 2020] and for manipulative marketing.

It is of interest that ethical issues in all of these areas were highlighted by Weizenbaum [1976] in his critique of the early AI "psychotherapist" ELIZA, which he himself had created. Weizenbaum argued that humans have no cognitive defenses against social AI, an argument that is currently played out in the international news.

7.2. *Could embodied AIs be ethical agents?*

The question of embodiment arose during the war over architecture as discussed in §3.2 above, but did not achieve practical importance until the robotics revolution of the early 2000s. The issue of embodiment goes, however, beyond architecture to debates about metaphysics and ontology, and offers surprising insights into the debates about ethics, particularly the possibly of AIs as ethical agents, within the ongoing AI wars.

The contemporary caricature of AI ethics comes in the form of Asimovs three (four) laws of robotics. Of course, no one who has read any of Asimov's stories would be tempted to take these seriously, as the "laws" feature in the stories precisely to highlight the many surprising and easy ways they can be violated. Yet certain attempts to talk about AI ethics start at roughly the same place, with simple, rule-driven systems [Bringsjord & Taylor, 2012; Bello & Bringsjord, 2013]. Most of these case studies seem to approach the problem exactly as Asimov imagined in his fiction, varying in what the exact rules are, or whether they follow something like a deontological system or a utilitarian one. But the questions that drive the current

and future ethical AI Wars can be seen growing directly out of those previous debates, largely nesting in the places where questions remained unsolved even as other questions seemed to reach resolution.

When philosophers talk about embodiment in relation to AI, everyone is quick to point out that a computer is a physical system in a real space, and is thus embodied, in a sense. While this is true, the real force of the embodiment debates is more about the capacities of materials and form of embodiment a machine may take in relation to the kind of mind that researchers hope will result [Anderson, 2003]. There are a number of ways this matters when we think about the future of AI. First, there are social questions with ethical counterparts as discussed in §7.1 above: why are AI assistants (like Alexa and Siri) gendered as women [Siegel, Breazeal & Norton, 2009; Shaw-Garlock, 2014; Otterbacher & Talias, 2017]; why are so many humanoid corporate robots white [Bartneck *et al.*, 2018]; how does the form of the robot affect how humans relate to it, such as when companion social robots have animal forms like Paro, while concierge robots like Pepper have a diminutive humanish shape [Darling, 2017], and other questions of this sort. Second, there are questions about the influence that bodily forms have on concept-formation [Lakoff & Johnson, 1980, 1999] and, from there, it's a short trip to moral imagination [Johnson, 1993; Dewey, 1998; Fesmire, 2003; Brown, 2020] and hence, ethics. These questions relate back to the discussion of kinds of minds in §6, but take on increasing importance as we think about how this relates to questions about the conditions for the possibility of ethical AI. Finally, there are questions that relate embodiment to ethics more directly, through an enactive sense that marries consciousness to embodiment before enabling the possibility of ethics at all [Varela, 1999; Thompson, 2001; Torrance, 2008; Colombetti & Torrance, 2009; Torrance, 2014].

In a sense, everything old is new again in AI. The arguments about AI ethics about which rules are the correct rules, without acknowledging the problems in rule-based approaches generally, see their most promising rebuttal in the form of something like practical wisdom, or artificial phrōnesis [Gerdes, 2016; Sullins, 2016; Vallor, 2016]. Similar arguments can be found in Dewey [1998] and more contemporary versions in Varela [1999]. For example, Varela's discussion of the failed strict computationalist paradigm in AI points out that "early optimism has given way to the recent and growing conviction that artificial intelligence worthy of the name will not be achieved without first understanding the situated embodiments of simple acts" (p. 8). And while some contemporary theorists reconcile an embodied, situated view with a version of computationalism (again see §3.2), many remain unsure such a reconciliation is possible [e.g. Di Paolo, 2003]. The problem, as Varela [1999] puts it, is "that the cognitive structures of human life emerge from recurrent sensorimotor patterns" (p. 15) and so it is the sensorimotor patterns that AI researchers are trying to replicate, but often without anything like a body with an (analog for an) appropriate kind of sensorimotor system.

Whether we're thinking about ethics or more mundane activities in AI, criticisms that emphasize embodiment raise the Frame Problem (see §5) in a new form.

Humans and other organisms are always embedded in some specific circumstance, are always ready for action, and move quickly from one state of readiness-for-action to another. Embodied robots must do the same. The Frame Problem can be viewed as an infinite regress, demanding rules for the application of other rules all the way down. Here the claim is that any such regress hits a background of sensorimotor habit, or at an even lower level, structural or architectural capability. In the face of an unsolvable Frame Problem, all any system can do is cope: guess and hope for the best. Artificial phrōnesis – skillful coping – becomes a potential way forward.

8. Conclusions

Philosophers do not solve problems [Dietrich, 2011]; their job is to ask questions. They have been asking questions about knowledge, inference, rationality, and morality for millenia. It is not, therefore, surprising that philosophers have had a difficult time with AI, with its move-fast-and-break-things culture, its enthusiasms, and its rapid penetration into every aspect of ordinary life. Neither is it surprising that many – quite possibly most – in AI have little patience for philosophers. If your job is to build something that works, nagging subtleties are not your first concern.

Great Philosophical Objections to Artificial Intelligence: The History and Legacy of the AI Wars traces the history of this difficult relationship from Turing [1950] to the present, showing how the first wave of philosophical attacks on AI morphed into the present, more nuanced and multidisciplinary, discussions of consciousness and ethics. While popular culture continues to offer a steady stream of fantasy AIs that look and act like humans, philosophical discussions of AI increasing concern actual or near-future AI systems for which being an "artificial human" is not the goal. As with industrial robots, most of these replicate and extend only some task-specific fraction of human cognition; autonomous vehicles, securities-trading systems, and even sex robots are valuable precisely because they do *not* replicate human cognition or behavior in general. Experimental platforms like the iCub are mini-AGIs, but are explicitly tools, not artificial children.

We can therefore ask: has the original goal – assuming it was a goal – of human-like AI become obsolete? Are we not in an era of surpassing human capabilities in narrow areas, while trying to avoid the pitfalls of human cognition in general? As our understanding of the Frame Problem has improved, we have come to see more clearly how far humans fall short of solving it. From Simon [1972] onward, studies of human cognition in the laboratory and in the field have demonstrated that the ideal rational agent of Aristotle, Descartes, or John Stuart Mill is not even a good approximation. As AI, driven by opportunity or necessity, turns more to the task of constructing distinctly *non*-human forms of intelligence, how will philosophy respond? The AI Wars, we expect, will only get more interesting.

Acknowledgments

We thank the computers everywhere who labored silently for us.

References

Aaronson, S. [2014a] My Conversation with "Eugene Goostman," the chatbot that's all over the news for allegedly passing the Turing test. https://www.scottaaronson.com/blog/?p=1858

Aaronson, S. [2014b] Why I am not an integrated information theorist (or, the unconscious expander). https://www.scottaaronson.com/blog/?p=1799

Adam, A. [1998] *Artificial Knowing: Gender and the Thinking Machine* (Routledge).

Anderson, M. L. [2003] Embodied cognition: A field guide. *Artif. Intell.* **149**, 91–130.

Anderson, M. & Anderson, S. [2011] *Machine Ethics* (Cambridge University Press).

Anderson, J. M., Nidhi, K., Stanley, K. D., Sorensen, P., Samaras, C. & Oluwatola, O. A. [2014] Autonomous vehicle technology: A guide for policymakers. Rand Corporation. https://www.rand.org/pubs/research_reports/RR443-2.html

Arkin, R. C. [2008] Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction* pp. 121–128 https://doi.org/10.1145/1349822.1349839.

Arkin R. C. [2009] *Governing Lethal Behavior in Autonomous Robots* (Chapman & Hall).

Arkin, R. [2015] The case for banning killer robots: Counterpoint. *Comms. ACM* **58**(12), 46–47.

Arquilla, J. [2010] The new rules of war. *Forign Policy* http://www.foreignpolicy.com/articles/2010/02/22/the_new_rules_of_war

Asaro, P. [2012] On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *Int. Rev. Red Cross* **94**(886), 687–709.

Asimov, I. [1950] *I, Robot* (Gnome, New York).

Bargh, J. A. & Ferguson, M. J. [2000] Beyond behaviorism: On the automaticity of higher mental processes. *Psychol. Bull.* **126**, 925–945.

Bargh, J. A., Schwader, K. L., Hailey, S. E., Dyer, R. L. & Boothby, E. J. [2012] Automaticity in social-cognitive processes. *Trends Cogn. Sci.* **16**, 593–605.

Bartneck, C., Yogeeswaran, K., Ser, Q. M., Woodward, G., Wang, S., Sparrow, R. & Eyssel, F. [2018] Robots and racism. *Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction* (HRI 18) ACM, pp. 196–204.

Bartoletti, I. [2020] *An Artificial Revolution: On Power, Politics and AI* (Chicago: Indigo).

Bello, P. & Bringsjord, S. [2013] On how to build a moral machine. *Topoi* **32**(2), 251–266.

Benjamin, R. [2019] Race after technology: Abolitionist tools for the new Jim code. *Socoial Forces* **98**(4), 1–3.

Bhargava, V. & Kim, T. W. [2017] Autonomous vehicles and moral uncertainty, in: P. Lin, K Abney & R. Jenkins (eds) *Robot Ethics, 2.0: From Autonomous Cars to Artificial Intelligence* (MIT) pp. 5–20.

Boly, M., Sanders, R. D., Mashour, G. A. & Laureys, S. [2013] Consciousness and responsiveness: Lessons from anaesthesia and the vegetative state, *Curr. Opin. Anesth.* **26**(4), 444–449.

Bonnefon, J. F., Shariff, A. & Rahwan, I. [2016] The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576.

Bonnemains, V., Saurel, C. & Tessier, C. [2018] Embedded ethics: Some technical and ethical challenges. *Ethics Inform. Tech.* **20**(1), 41–58.

Bostrom, N. [2003] Ethical issues in advanced artificial intelligence, in: S. Schneider (ed.) *Science Fiction and Philosophy: From Time Travel to Superintelligence* (Wiley) 277–284.

Bringsjord, S. & Taylor, J. [2012] The divine-command approach to robot ethics, in: P. Lin, K. Abney & G. Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press) pp. 85–108.

Brooks, R. [1991] Intelligence without representation. *Artif. Intell.* **47**, 139–159.

Brooks, R. A. [2003] *Flesh and Machines: How Robots Will Change Us* (Vintage).

Brown, M. [2020] *Science and Moral Imagination: A New Ideal for Values in Science* (University of Pittsburgh Press).

Bunge, M. [1977] Towards a technoethics. *Monist* **60**(1), 96–107.

Calvo, P., Baluška, F. & Trewavas A. [2020] Integrated information as a possible basis for plant consciousness. Biochem. Biophys. Res. Comm. in press. `https://doi.org/10.1016/j.bbrc.2020.10.022`

Carter, O., Hohwy, J., van Boxtel, J., Lamme, V., Block, N., Koch, C. & Tsuchiya, N. [2018] Conscious machines: Defining questions. *Science* **359**, 400.

Chalmers, D. J. [1996] *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press).

Champagne, M. & Tonkens, R. [2015] Bridging the responsibility gap in automated warfare. *Phil. Technol.* **28**(1), 125–137.

Chater N. [2018] *The Mind is Flat* (Allen Lane).

Chomsky, N. [1959] Review of B. F. Skinner, *Verbal Behavior*. *Language* **35**, 26–58.

Church, A. [1936] An unsolvable problem of elementary number theory. *Am. J. Math.* **58**, 345–363.

Cianciolo, A. T., Matthew, C., Sternberg, R. J. & Wagner, R. K. [2006] Tacit knowledge, practical intelligence, and expertise, in: K. A. Ericsson, N. Charnes, R. R. Hoffman & P. J. Feltovich (eds.), *Cambridge Hnadbook of Expertise and Expert Performance* (Cambridge University Press) pp. 613–632.

Coeckelbergh, M. [2020] *AI Ethics* (MIT Press).

Colombetti, G. & Torrance, S. [2009] Emotion and ethics: An inter-(en)active approach. *Phenom. Cogn. Sci.* **8**, 505–526.

Dabringer G. (ed.) [2011] *Ethical and Legal Aspects of Unmanned Systems* (Vienna: Insti-
tut für Religion und Frieden) `https://www.researchgate.net/publication/313114261_Ethical_and_Legal_Aspects_of_Unmanned_Systems_Interviews`

Danaher, J. & McArthur, N. [2017] *Robot Sex* (MIT Press).

Darling, K. [2017]. Whos Johnny: Anthropomorphic framing in human-robot interaction, integration, and policy, in: P. Lin, K. Abney & R. Jenkins (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford University Press) pp. 173–188.

De Sio, F. S. [2017] Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theor. Moral Pract.* **20**(2), 411–429.

Dehaene, S., Lau, H. & Kouider, S. [2017] What is consciousness, and could machines have it? *Science* **358**, 486–492.

Dennett, D. [1987] *The Intentional Stance* (MIT/Bradford).

Dennett, D. C.]1998] When HAL kills, whos to blame? in D. G. Stork (ed.) *HAL's Legacy: 2001's Computer as Dream and Reality* (MIT Press) pp. 351–365.

Dewey, J. [1998] Evolution and ethics, in: L. Hickman & T. Alexander (eds.) *The Essential Dewey*, Volume 2. (Indiana University Press) pp. 225–236.

Dietrtich, E. [2011] There is no progress in philosophy. *Essays Phil.* **12**(2), 330–345.

Dietrich, E. & Fields, C. [2020] Equivalence of the Frame and Halting Problems. *Algorithms* **13**, 175.

Dietrich, E., Fields, C., Sullins, J., van Heuveln, B. & Zebrowski, R. [2021] *Great Philosophical Objections to Artificial Intelligence: The history and Legacy of the AI Wars* (Bloomsbury Academic).

Dignum, V. [2019] *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Springer).

Di Paolo, E. A. [2003] Organismically-inspired robotics: Homeostatic adaptation and teleology beyond the closed sensorimotor loop, in: K. Murase & T. Asakura (eds.) *Dynamical Systems Approach to Embodiment and Sociality* (Magill, South Australia: Advanced Knowledge International) pp 19–42.

Docherty, B. L. [2015] *Mind the gap: The Lack of Accountability for Killer Robots* (Human Rights Watch) `https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots#`

Doerig, A., Schurger, A., Hess, K. & Herzog, M. H. [2019] The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Cons. Cogn.* **72**, 49–59.

Doubleday, J. [2019] Pentagon to grapple with tough questions about using AI for warfare. *Inside the Pentagon's Inside Missile Defense* **25**(23).

Dreyfus, H. [1972] *What Computers Can't Do* (MIT Press).

Drozdek, A.]1992] Moral dimension of man and artificial intelligence. *AI & Society* **6**(3), 271–280.

Dummett, M. [1978] *Truth and Other Enigmas* (Harvard University Press).

Etzioni, A. & Etzioni, O. [2017] Incorporating ethics into artificial intelligence. *J. Ethics* **21**(4), 403–418.

Fesmire, S. [2003] *John Dewey and Moral Imagination: Pragmatism in Ethics* (Indiana University Press).

Fleetwood, J. [2017] Public health, ethics, and autonomous vehicles. *Am. J. Public Health* **107**(4), 532–537.

Floridi, L. [2008] Information ethics: Its nature and scope, in: J. Van Den Hoven & J. Weckert (eds.) *Information Technology and Moral Philosophy* (Cambridge University Press) pp. 40–65.

Floridi, L. & Chiriatti, M. [2020] GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **30**, 681–694.

Floridi, L. & Taddeo, M. [2014] *The Ethics of Information Warfare* (Vol. 14, Law, Governance and Technology series) (Springer). .

Fodor, J. [1975] *The Language of Thought* (Harvard University Press).

Fodor, J. [1987] Modules, frames, fridgeons, sleeping dogs, and the music of the spheres, in: Z. Pylyshyn (ed.) *The Robots Dilemma: The Frame Problem in Artificial Intelligence* (Ablex) pp. 139–149.

Foot, P. [1967/2002] The problem of abortion and the doctrine of double effect, in: *Virtues and Vices* (Oxford University Press) pp.10–17.

Franklin, S., Madl, T., D'Mello, S. & Snaider, J. [2014] LIDA: a systems-level architecture for cognition, emotion and learning. *IEEE Trans. Auton. Mental Devel.* **6**, 19–41.

Frege, G. [1892] Über Sinn und Bedeutung (On sense and reference). *Z. Phil. Philos. Kritik* **100**, 25–50.

Gagliano, M. & Grimonprez, M. [2015] Breaking the silenceLanguage and the making of meaning in plants. *Ecopsychol.* **7**, 145–152.

Ghaffary, S. [2021] Googles new union, briefly explained: The new Alphabet Workers Union is an important step forward, but it has a long road ahead. *Vox, Recode* `https://www.vox.com/recode/22213494/google-union-alphabet-workers-tech-organizing-activism-labor`

Gerdes, A. [2016] The role of phrōnesis in robot ethics, in: J. Seibt, M. Noskov & S. S. Andersen (eds.) *What Social Robots Can and Should Do* (IOS Press) pp. 129–135.

Gerdes J. & Thornton S. [2015] Implementable ethics for autonomous vehicles, in: M. Maurer, J. Gerdes, B. Lenz & H. Winner (eds.) *Autonomes Fahrenn: Technische, rechtliche und gesellschaftliche Aspekte* (Springer) pp. 87–102.

Gibson, J. J. [1979] *The Ecological Approach to Visual Perception* (Houghton Mifflin).

Gödel, K. [1931] Über formal unentscheidbare sätze der *Principia Mathematica* und verwandter systeme, i. *Monatshefte für Mathematik und Physik* **38**(1), 173–198.

Goertzel, B. [2014] Artificial general intelligence: Concept, state of the art, and future prospects. *J. Artif. Gen. Intell.* **5**, 1–46.

Goff, P. [2019] *Galileo's Error: Foundations for a New Science of Consciousness* (Vintage).

Goodall, N. J. [2014] Machine ethics and automated vehicles, in: G. Meyer & S. Beiker (eds) *Road Vehicle Automation 4* (Springer) pp. 93–102.

Goodall, N. J. [2016] Away from trolley problems and toward risk management. *Appl. Artif. Intell.* **30**(8), 810–821.

Govindarajulu, N. S. & Bringsjord, S. [2017] On automating the doctrine of double effect. Preprint arXiv:1703.08922.

Griggs, T. [2018] How a self-driving Uber killed a pedestrian in Arizona. *New York Times* `https://www.nytimes.com/interactive/2018/03/20/us/self-drivinguber-pedestriankilled`

Gunkel, D. J. [2017] Mind the gap: responsible robotics and the problem of responsibility. *Ethics Inform Tech.* **22**, 307–320.

Güntürkün, O. & Bugnyar, T. [2016] Cognition without cortex. *Trends Cogn. Sci.* **20**, 291–303.

Gurney, J. K. [2015] Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Albuq. Law Rev.* **79**, 183–267.

Hall, J. S. [2009] *Beyond AI: Creating the Conscience of the Machine* (Prometheus).

Hawkes, R. & Lacey, C. [2019] The future of sex: Intermedial desire between fembot fantasies and sexbot technologies. *J. Popular Culture* **52**(1), 98–116.

Hayes, P. [1987] What the frame problem is and isnt, in: Z. Pylyshyn (ed.) *The Robots Dilemma: The Frame Problem in Artificial Intelligence* (Ablex) pp. 123–137.

Himmelreich, J. [2018] Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theor. Moral Pract.* **21**(3), 669–684.

Howard, B. [2015] Musk, Hawking, Wozniak: Ban AI warfare, autonomous weapons. ExtremeTech.com, 2015-07-27. `https://www.extremetech.com/extreme/210983-musk-hawking-woz-ban-ai-warfare-autonomous-weapons`

Jodi R. [2015] Leaderless Palestinian youth, inspired by social media, drive rise in violence in Israel. *New York Times* `https://www.nytimes.com/2015/10/14/world/middleeast/leaderless-palestinian-youth-inspired-by-social-media-drive-a-rise-in-violence.html`

Johnson, D. G. [1985/2001] *Computer Ethics*, 3rd Ed. (Prentice-Hall).

Johnson, M. [1993] *Moral Imagination: Implications of Cognitive Science for Ethics* (University of Chicago Press).

Kleiner, J. & Hoel, E. [2020] Falsification and consciousness. `https://arxiv.org/abs/2004.03541`

Kowert, W.]2017] The foreseeability of human-artificial intelligence interactions. *Texas Law Rev.* **96**(1), 181–204.

Kubes, T. [2019] New materialist perspectives on sex robots: A feminist dystopia/utopia? *Social Sciences* (Basel) **8**(8), 224 –238.

Lakoff, G. & Johnson, M. [1980] *Metaphors We Live By* (University of Chicago Press).

Lakoff, G. & Johnson, M. [1999] *Philosophy in the Flesh: The Embodied Mind and*

its Challenge to Western Thought (Basic Books).

LeCun, Y., Bengio, Y. & Hinton, G. [2015] Deep learning. *Nature* **521**, 436–444.

Lenat, D. & Brown, J. S. [1983] Why AM and Eurisko appear to work, in *Proc. AAAI-83* (AAAI), pp. 236–240.

Leong, B. [2019] Facial recognition and the future of privacy: I always feel like ... somebodys watching me. *Bull. Atomic Sci.* **75**(3), 109–115.

Levin M. [2019] The computational boundary of a self: Developmental bioelectricity drives multicellularity and scale-free cognition. *Front. Psychol.* **10** 2688.

Levy, D. [2009] *Love and Sex with Robots: The Evolution of Human-Robot Relationships* (Harper Perennial).

Lighthill, J. [1973] Artificial Intelligence: A general survey, in *Artificial Intelligence: A Paper Symposium* (London, Science Research Council), pp. 1–21.

Lin, P.]2016] Why ethics matters for autonomous cars, in: M. Maurer, B. Lenz & J. C. Gerdes (eds.) *Autonomous Driving* (Springer) pp. 69–85.

Lin, P. [2017] Robot cars and fake ethical dilemmas. *Forbes* `https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethicaldilemmas/#53bd0e2213a2`

Lin, P., Abney, K. & Bekey, G. A. [2011] *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press).

Lin, P., Abney, K. & Jenkins, R. [2017a] *Robot Ethics, 2.0: From Autonomous Cars to Artificial Intelligence* (MIT Press).

Lin, P., Bekey, G. & Abney, K. [2008] *Autonomous Military Robotics: Risk, Ethics, and Design* (San Luis Obispo: California Polytechnic State University) `http://ethics.calpoly.edu/ONR_report.pdf`

Lin, P., Mehlman, M., Abney, K. & Galliott, J. [2014] Super soldiers (Part 1): What is military human enhancement? in: S. Thompson & J. Thompson (eds.) *Global Issues and Ethical Considerations in Human Enhancement Technologies* (IGI Global) pp. 119–138.

Lucas, J. R. [1961] Minds, machines, and Gödel. *Philosophy* **XXXVI**, 112–127.

Lucas, J. R. [1996] Minds, machines, and Gödel, a retrospect, in P. J. R. Millican & A. Clark (eds.), Machines and Thought: The Legacy of Alan Turing, Vol. 1 (Oxford University Press), pp. 103–124.

Lyon P. [2020] Of what is "minimal cognition the half-baked version? *Adapt. Behav.* **28**, 407–428.

Madison, G. [1991] Merleau-Pontys deconstruction of logocentrism, in M. Dillon (ed.), *Merleau-Ponty Vivant* (SUNY Press) pp. 117–152.

Mather J. [2019] What is in an octopuss mind? *Animal Sent.* **4**, 26(1).

Maturana, H. R. & Varela, F. J. [1980] *Autopoesis and Cognition: The Realization of the Living* (D. Reidel).

McCarthy, J. and Hayes, P. [1969] Some philosophical problems from the standpoint of artificial intelligence, in B. Meltzer & D. Michie (eds.), *Machine Intelligence*, Vol 4. (Edinburgh University Press), pp. 463–502.

McCulloch, W. S. & Pitts, W. [1943] A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133.

Minsky, M. L. and Papert, S. A. [1969] *Perceptrons* (MIT Press).

Moor, J. H. [2011] The nature, importance, and difficulty of machine ethics, in: M. Anderson & S. Anderson (eds.) *Machine Ethics* (Cambridge University Press) pp. 13–20.

Newell, A. [1980] Physical symbol systems. *Cogn. Sci.* **4**, 135–183.

Newell, A. [1990] *Unified Theories of Cognition* (Harvard University Press).

Newell, A., Shaw, J. C. & Simon, H. A. [1957] Empirical explorations in the Logic Theory machine, in M. M. Astrahan (ed.), *Proc. 1957 Western Joint Computer Conference* (ACM, New York), pp. 218–230.

Newell, A. & Simon, H. A. [1961] GPS: A program that simulates human thought, in H. Billing (ed.), *Lernende Automaten* (Oldenbourg, Munich), pp. 109–124.

Newell, A. & Simon, H. A. [1976] Computer science as empirical inquiry: Symbols and search. *Comm. ACM* **9**(3), 113–126.

Nieder, A., Wagener, L. & Rinnert, P. [2020] A neural correlate of sensory consciousness in a corvid bird. *Science* **369**, 1626–1629.

Noble, S. [2018] *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press).

Oizumi, M., Albantakis, L. & Tononi, G. [2014] From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comp. Biol.* **10**, e1003588.

Otterbacher, J. and Talias, M. [2017] S/He's too warm/agentic!: The influence of gender on uncanny reactions to robots. *Proceedings of the 2017 ACM/IEEE International conference on HRI* pp. 214–223 `https://doi.org/10.1145/2909824.3020220`

Oudeyer, P.-Y. & Kaplan, F. [2007] What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* **1**, 6.

Panksepp J. [2005] Affective consciousness: Core emotional feelings in animals and humans. *Cons. Cognit.* **14**, 30–80.

Peeters, A. & Haselager, P. [2019] Designing virtuous sex robots. *Int. J. Social Robotics* in press `https://link.springer.com/article/10.1007/s12369-019-00592-1`

Penrose, R. [1989] *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford University Press).

Penrose, R. [1994] *Shadows of the Mind* (Oxford University Press).

Picard, R. W. [2010] Affective computing: From laughter to IEEE. *IEEE Trans. Affect. Comp.* **1**(1), 11–17.

Reardon, S. [2019] Rival theories face off over brain's source of consciousness. *Science* **366**, 293.

Rees, G., Kreiman, G. & Koch, C. [2002] Neural correlates of consciousness in humans. *Nat. Rev. Neurosci.* **3**, 261–270.

Richardson, K. [2016] The asymmetrical 'relationship'. *Computers & Society* **45**(3), 290–293.

Rochat P. [2003] Primordial sense of embodied self-unity, in: V. Slaughter & C. A. Brownell (eds.), *Early Development of Body Representations.* (Cambridge University Press), pp. 3–18.

Roth, C., Maziéres, A. & Menezes, T. [2020] Tubes and bubbles: Topological confinement of YouTube recommendations. *PloS One* **15**(4), e0231703.

Rumelhart, D. E., McClelland, J. L. & PDP Research Group [1986] *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (MIT Press).

Schank, R. & Abelson, R. [1977] *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures* (Lawrence Erlbaum).

Schnell, A. K., Amodio, P., Boeckle, M. & Clayton, N. S. [2020] How intelligent is a cephalopod? Lessons from comparative cognition. *Biol. Rev.* in press. `https://doi.org/10.1111/brv.12651`

Searle, J. R. [1980] Minds, brains, and programs, *Behav. Brain Sci.* **3**(3), 417–424.

Senior, A. W., Evans, R., Jumper J. *et al.* [2020] Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710.

Sharkey, M. & Sharkey, A [2019] Autonomous weapons systems, killer robots and human dignity. *Ethics Inform Tech.* **21**(2), 75–87.

Shaw-Garlock, G. [2014] Gendered by design: Gender codes in social robotics, in: M. Noskov (ed.) *Social Robots: Boundaries, Potential, Challenges* (Routledge) pp. 199–218.

Siegel, M., Breazeal, C. & Norton, M. I. [2009] Persuasive robotics: The influence of robot gender on human behavior. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* pp. 2563–2568 `https://doi.org/10.1109/IROS.2009.5354116`

Simon, H. A. [1972] Theories of bounded rationality, in: C. B. McGuire & R. Radner (eds.) *Decision and Organization* (North-Holland) pp. 161–176.

Singer, P. [2009] *Wired for War: The Robotics Revolution and Conflict in the Twenty-First Century* (Penguin).

Sloman, A. [1978] *The Computer Revolution in Philosophy: Philosophy, Science, and Models of the Mind* (Harvester).

Smolensky, P. [1988] On the proper treatment of connectionism. *Behav. Brain Sci.* **11**, 1–23.

Strawson, G. [2006] *Consciousness and its place in nature: does physicalism entail panpsychism?* (Imprint Academic).

Suddendorf, T. & Corballis, M. C. [2007] The evolution of foresight: What is mental time travel and is it unique to humans? *Behav. Brain Sci.* **30**, 299–313.

Sullins, J. P. [2005] Ethics and artificial life: From modeling to moral agents. *Ethics Inform Tech.* **7**(3), 139–148.

Sullins, J. P. [2010] RoboWarfare: Can robots be more ethical than humans on the

battlefield? *Ethics Inform Tech.* **12**(3), 263–275.

Sullins, J. P. [2012] Robots, love, and sex: The ethics of building a love machine. *IEEE Trans. Affect. Comput.* **3**(4), 398–409.

Sullins, J. P. [2016] Artificial phrōnesis and the social robot, in: J. Seibt *et al.* (eds.) *What Social Robots Can and Should Do* (IOS Press) pp. 37–39.

Sun, R. [2007] The importance of cognitive architectures: An analysis based on CLARION. *J. Expt. Theor. Artif. Intel.* **19**, 159–193.

Taylor, J. E. T. & Taylor, G. W. [2020] Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychon. Bull. Rev.* in press. `https://doi.org/10.3758/s13423-020-01825-5`

Thompson, E. [2001] Empathy and consciousness. *J. Cons. Stud.* **8**(5), 1–32.

Tononi, G. [2014] Why Scott should stare at a blank wall and reconsider (or, the conscious grid). `https://www.scottaaronson.com/blog/?p=1823`

Tononi, G. & Koch, C. [2015] Consciousness: Here, there and everywhere? *Phil. Trans. R. Soc. B* **370**, 20140167.

Torrance, S. [2008] Ethics and consciousness in artificial agents. *AI & Society* **22**, 495–521.

Torrance, S. [2014] Artificial consciousness and artificial ethics: Between realism and social relationism. *Philos. Technol.* **27**, 9–29.

Turing, A. M. [1936] On computable numbers, with an application to the *Entscheidungsproblem*. Proc. London Math. Soc. Ser. 2 **42**, 230–265.

Turing, A. M. [1948] Intelligence machinery. Report to the National Physical Laboratory, UK. Reprinted in Ince, D. C. (ed.) [1992] *Mechanical Intelligence* Vol. I. (Elsevier) pp. 107–127.

Turing, A. M. [1950] Computing machines and intelligence, *Mind* **59**(236), 433–460.

Turing, A. M. [1952] The chemical basis of morphogenesis. *Phil. Trans. R. Soc. B* **237**, 37–72.

Vallor, S. [2016] *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press).

van Gelder, T. [1998] The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* **21**, 637–638.

Varela, F. [1999]. *Ethical Know-How: Action, Wisdom, and Cognition* (Stanford University Press).

Veruggio, G. & Operto, F. [2008] Roboethics: Social and ethical implications of robotics. *Springer Handbook of Robotics* (Springer) pp. 1499-1524.

Wallach, W. & Allen, C. [2008] *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press).

Wallach, W. & Allen, C. [2013] Framing robot arms control. *Ethics Inform. Tech.* **15**(2), 125–135.

Weizenbaum, J. [1976] *Computer Power and Human Reason: From Judgment to Calculation* (Freeman).

Zuckerkandl, E. & Pauling, L. [1965] Evolutionary divergence and convergence in

proteins, in V. Bryson & H. J. Vogel (eds.), *Evolving Genes and Proteins* (Academic Press) pp. 97–166.