Zebrowski, Robin

Artificial instinct : Lem's robots as a model case for AI

Pro-Fil. 2021, vol. 22, iss. Special issue, pp. 92-102

ISSN 1212-9097 (online)

Stable URL (DOI): https://doi.org/10.5817/pf21-3-2423
Stable URL (handle): https://hdl.handle.net/11222.digilib/144853

Pro-Fil (Special Issue) 2021: 92–102
https://doi.org/10.5817/pf21-3-2423

ARTIFICIAL INSTINCT: LEM'S ROBOTS AS A MODEL CASE FOR AI

ROBIN ZEBROWSKI

Departments of Philosophy, Psychology, and Computer Science, Beloit College,
Beloit, WI, USA, zebrowsr@beloit.edu

RESEARCH PAPER ▪ SUBMITTED: 17/10/2021 ▪ ACCEPTED: 1/12/2021

Abstract: In the seventy years since AI became a field of study, the theoretical work of philosophers has played increasingly important roles in understanding many aspects of the AI project, from the metaphysics of mind and what kinds of systems can or cannot implement them, the epistemology of objectivity and algorithmic bias, the ethics of automation, drones, and specific implementations of AI, as well as analyses of AI embedded in social contexts (for example). Serious scholarship in AI ethics sometimes quotes Asimov's speculative laws of robotics as if they were genuine proposals, and yet Lem remains historically undervalued as a theorist who uses fiction as his vehicle. Here, I argue that Lem's fiction (in particular his fiction about robots) is overlooked but highly nuanced philosophy of AI, and that we should recognize the lessons he tried to offer us, which focused on the human and social failures rather than technological breakdowns. Stories like "How the World Was Saved" and "Upside Down Evolution" ask serious philosophical questions about AI metaphysics and ethics, and offer insightful answers that deserve more attention. Highlighting some of this work from *The Cyberiad* and the stories in *Mortal Engines* in particular, I argue that the time has never been more appropriate to attend to his philosophy in light of the widespread technological and social failures brought about by the quest for artificial intelligence. In service of this argument, I discuss some of the history and philosophical debates around AI in the last decades, as well as contemporary events that illustrate Lem's strongest claims in critique of the human side of AI.

Keywords: Stanislaw Lem; Asimov's Laws of Robotics; artificial intelligence

There have, of late, been many arguments that boil down to the idea that the problems in artificial intelligence (AI) and machine learning (ML) are caused or exacerbated by a lack of critical engagement with the humanities on the part of tech culture broadly. This can be seen through things like the breakdown of Google's external AI Ethics council before it even began (D'Onfro 2019), high-profile firings of ethics researchers at Google (BBC News February 2021; BBC News December 2020), and overwhelming pushes in education toward STEM disciplines at the expense of the humanities (Shamir 2020, Zakaria 2015). Examples of such include claims that AI discourse focuses too much on non-problems (like robot rights) and not enough focused on actually-existing current problems (like human rights) (Birhane – van Dijk 2020) or that people working in AI tend to ignore anticipated negative or societal outcomes of their work, if bothering to look for them at all (Birhane et al. 2021). These arguments are not new, and philosophers have been trying to engage AI researchers and roboticists in thinking about the ethics surrounding their work for a long time (Sullins 2015).

There are two pieces to the argument I will make here. First, I will argue that the metaphysics/ontology of AI as Lem understood them through his robot stories[1] offer a richly-supported picture of beings flawed in the ways that humans are flawed; and because human-like minds are the explicit goal of the AI project, it is problematic that AI is generally pursued without a recognition or acknowledgment of the many ways human cognition is flawed by design (a feature often emphasized humanities-based pursuits like fiction, incidentally). When I say "flawed by design," I mean ways that our cognitive functions can go wrong in spite of, and likely because of, the way they work. You can think here about optical illusions, cognitive heuristics, and even the way emotion guides reasoning (Damasio 1994). Indeed, once algorithmic AI became a matter staked out as strictly for computer scientists, edging out philosophers and cognitive scientists and their expertise in human cognition, many of the current problems in AI began. Builders of AI claim to want human-like intelligence, but what the real goal appears to be is more akin to a magic rationality machine. Lem not only satirizes such magic machines, but also shows us what we'd really be getting with human-like AI as our goal. Had researchers read and considered Lem's work earlier, perhaps we would've reconceived the sorts of systems we ought to be trying to create, or at least been prepared for the exact problems we now routinely see with these systems. Additionally, using AI Ethics as a kind of case study, I will show that, while it has long been normalized to have some engagement with fiction as a way of imagining our possible futures, the choice on behalf of AI researchers and theorists to use Asimov's Laws of Robotics has consistently narrowed our understanding of the problems rather than expanded the possibilities or clarified the systems in the ethics of AI. If we look to Stanislaw Lem's robot stories, however, we find keen insight into many of the problems in AI, including AI ethics, but also the ontological underpinnings of AI theory itself.

First, a terminological distinction must be made. The phrase "artificial intelligence" (AI) has come to mean at least two very different things, both to researchers and the general public; on the one hand, it is shorthand for the algorithms trained on large datasets, often using some sort of machine learning or deep learning technique, often deployed as expert systems in a single domain. On the other hand, it has meant an almost mad-science sort of pursuit of building a machine that not just mimics human thought and behavior, but experiences the world and has a human-like mind. This is sometimes referred to as AGI (artificial general intelligence) or Strong AI. The earliest AI research makes no clear distinction between these projects, however (Weizenbaum 1966). The idea behind the strong version of the AI project is not to simulate a mind (either to study it or to deploy it as an expert system) but instead to literally create a mind. There have always been multiple methodologies engaged in this project, not always compatible with one another (the traditional GOFAI approach of high-level symbolic processing, artificial neural nets/connectionism, and cognitive or embodied robotics as the primary research agendas). Lem has satirized them all, and has understood them at a level many practicing researchers still do not.

The most classic definition of the GOFAI ("Good Old Fashioned AI") approach comes from Newell and Simon's famous Turing Award lecture, where they formulate the Physical Symbol System Hypothesis: "A physical symbol system has the necessary and sufficient means for general intelligent action. By 'necessary' we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By 'sufficient' we mean that any physical symbol

[1] I am only using Lem's fiction here intentionally, as most of his non-fiction work has not been widely available in English translation for nearly as long as his robot stories have.

system of sufficient size can be organized further to exhibit general intelligence. *By 'general intelligent action' we wish to indicate the same scope of intelligence as we see in human action*…" (emphasis mine, 1976, 116). Human-level intelligence, then, has really been with us since the beginning.

Human-like AI

Looking back at the history of AI, the primary goal, both implicit and explicit, has been to replicate human intelligence and the human mind: (Dietrich et al. 2021a; Newell – Simon 1976; Turing 1950). We can look to something as simple as the Turing Test to see this: the controversial (and still unmet) benchmark in AI for seventy years has been the Turing Test, a test that is not just distinctly human-centric, but neurotypical human-centric. The test is so focused on distinctive human-like behavior that it's focused entirely on natural language, leaving out all kinds of human neurodiversity and intelligent non-human animals from participation. Of course, we are wise to remember that failing the test indicates nothing, as anyone one of us could fail to pass it simply by refusing to participate, or by the test being administered in another language (Dennett 1984). Much of Lem's robot fiction highlights this aspect of the quest for human-like AI, showing just the ways language-based tests are absurd, such as in "Trurl's Machine," "How the World Was Saved", and "The Tale of a Computer Who Fought a Dragon", each a vivid illustration of how the famed constructors (themselves robots) Trurl and Klapaucious battle with computers that play language games or miscalculate sums with absurdity (Lem 1976, 1992). In "The Tale of a Computer Who Fought a Dragon," for example, the world is saved by a hardware malfunction that changes the word "electrosaur" into "electrosauce" (Lem 1992).

There are real language-based worries in AI right now, and many of them grow from the ways algorithms are being trained up. The language-based NLP systems from OpenAI, GPT-2 and GPT-3, for example, have been shown to have radical limitations based on their training data. GPT-3 has been shown to exhibit a strong bias against Muslims (Abid – Farooqi – Zou 2021). Google's own internal research has shown that there are serious concerns with training these language models off of internet spaces, for reasons of representation, bias, and lack of accountability, among others (Bender et al. 2021). Lem's "How the World Was Saved," has always seemed a very intense critique of a very funny idea. In this story, the machine could unmake the world through language itself. But while Lem's machine merely unmade things that sound like nonsense to us now (implying that no trace of those concepts has been left behind), the power of big NLP algorithms to make entire people and ways of speaking and existing invisible is not nonsense; it's happening now.

There are folks who worry about "superintelligent AI" (Bostrom, 2014) which is supposed to be when AI intelligence surpasses human intelligence, but of course, depending on what domain we're talking about, that's already happened (a calculator, for example, far surpasses human speed and accuracy at mathematics, and my toaster is also way better than me at toasting bread). There's no reason to think any such superintelligent AI will ever exist in the general domain, and many such expert systems are already better than humans at a single task in a small domain. Lem's robot fiction, in a way, is one giant mockery of this idea, and rightfully so. Trurl and Klapaucious, for example, are always being hired to build various machines and robots for kings and rulers, and often those robots (or, more often than not, the kings themselves) go hilariously wrong in ways

both predictable and unpredictable. But part of what makes these stories so valuable is that our assumptions are always challenged when it comes to these sorts of machines. Lem rarely draws much attention to the fact that his robot constructors (those who construct robots) are also *robot* constructors (they themselves are robots, who do the constructing).

Unfounded fears about superintelligent AI aside, it has never been entirely clear why human intelligence should be the goal in AGI, except that we tend to measure all intelligence against human intelligence. The heavy reliance on linguistic capacities (seen, again, with the Turing Test, but also with the newest natural language models like GPT-3) puts this in stark relief. Wittgenstein famously wrote that if a lion could speak, we could not understand him (Wittgenstein 1953). But such insight has never stopped AI researchers from trying to train general intelligence into a system using nothing but language, ignoring the fact that our embodied practices and lifeworld are a big part of how we create and understand language. Lem's particular way of skewering this idea pits the two robot constructors against one another in "A Good Shellacking." We discover that Trurl has sent Klapaucious a robot to grant every wish- basically, not just superintelligent, but superpowered to the point of technological magic. This machine might remind us of a real-world counterpart that was promised to solve all kinds of problems, having almost endless capacity to be put to any task. I refer here to IBM's Watson computer, which had one very public success in beating the world champions at the game show Jeopardy. But when it was then sold to companies to solve other kinds of tasks, to be used in things like medical diagnostics and education, it was shown to be almost entirely useless (Lohr 2021). Not unlike Trurl's deception in "A Good Shellacking," where he himself is hiding in the machine pretending to be an extremely impressive machine in the style of the famed Mechanical Turk, Watson too ended up being a mostly-empty façade. In fact, many companies claiming to use AI to solve problems are just quietly using humans, but claiming it's AI, in the hopes, sometimes realized, of getting venture capital (Olson 2021).

Importantly, we need to ask what human-like AI would even look like. Once we do, we get a strange neurotypical picture of a flawed creature. Perhaps we imagine an ultimate rationality machine, like the Vulcans of Star Trek or HAL from 2001. But that isn't an idealized human, that's already a wildly different sort of system than humans (and other animals) are. Historically, cognitive science (and before it, philosophy) set up rationality as if it is opposed to emotion, but this is a false dichotomy, as much contemporary neuroscience (and a moment of reflection, unless perhaps you are Descartes) would reveal. Damasio, for example, shows that many decision-making processes only function optimally in humans when accompanied by and informed by an emotional response (1994). He outlines a somatic marker hypothesis in which our lived experiences build up bodily and neural shortcuts that we sometimes feel (like a gut feeling). These are built and honed over a lifetime, directly on top of evolutionarily older systems we all tend to be born with, and that function early as instinct, for example. When patients suffer specific kinds of frontal lobe damage, their somatic marker malfunctions and they can't always weigh their options very well as a result. We rely on all sorts of bodily, social, and cultural experiences to guide and develop our rationality. Emotions and feelings aren't opposed to that pure rationality, they're an enabling factor, but not one that most AI researchers have even considered. For humans, and likely for other animals as well, there is no such thing as pure reason, divorced from felt experience and emotion. Instead, our choices are rational insofar as we can weigh our values and expected outcomes of our actions. In his 1994 book, Damasio discusses a patient whose somatic marker system had been damaged. He tells a short anecdote of offering the patient two alternative dates for an upcoming appointment,

and how the patient was unable to make a simple decision. He says, "For the better part of a half-hour, the patient enumerated reasons for and against each of the two dates: previous engagements, proximity to other engagements, possible meteorological conditions, virtually anything that one could reasonably think about concerning a simple date… he was now walking us through a tiresome cost-benefit analysis, an endless outlining and fruitless comparison of options and possible consequences…" (193). Eventually, Damasio tells the patient to come on the second date, and the patient merely says, "that's fine." That 4e cognition (embodied, embedded, extended, and enactive) has been so slow to gain a foothold in AI speaks volumes about the assumptions in the field, and how disembodied and language-based they remain.

Unlike most robots in fiction, Lem's robots are, as his most famous translator says, "people too" (Kandel 1992). We see robots that are deceptive ("A Good Shellacking"), master poets ("Trurl's Electronic Bard"), mischievous ("The Mischief of King Balerion"), masters of information manipulation ("How Trurl Created a Demon of the Second Kind to Defeat the Pirate Pugg"), and, perhaps the largest and most enduring vice among all of Lem's robots, philosophers ("Tale of the Three Storytelling Machines of King Genius"). And why wouldn't we expect to see robots with vices and virtues, if our goal is human emulation? Humans are complex, and we have never been purely rational creatures. Trurl's Machine, the one that insisted that 2+2=7 and pulls itself out of its foundation to chase and hunt its constructor, is called a stupid machine for its troubles. But a person getting mad at being told they're wrong is an extremely human sight.

AI theorists need to either understand human cognition well enough to see that human-like AI is going to result in beings with all of these kinds of flaws (which are almost entirely features, not bugs of how our minds work), or they need to stop trying to build human-like AI. Lem showed us the all-too-human ways that robots might be, and while his stories are thought of as "fables for the cybernetic age," the moral they offer is extremely real: we're trying to build machines with speed and power we lack, but who are otherwise built the same way we are, designed with the same flaws evolution gave us, because that's the explicit goal of the mad-science kind of AI: human-like minds implemented in machines, functionally equivalent to humans in the relevant (cognitive) ways. Lem showed us exactly what we can expect from this endeavor.

AI Ethics (A case study)

It's difficult to read through scholarly literature in AI ethics and avoid seeing any mention of Asimov's three laws of robotics (I will refrain from listing them here, as their content is more or less irrelevant). It's just as difficult to avoid those laws in the literature where philosophy as a discipline engages with science fiction intentionally, with pedagogical purpose. In both cases, Asimov's laws are a launching point, a starting place to open up conversations about robots in society, and how we might go about programming ethical codes into machines so that they cannot misbehave. For example, the text now considered the benchmark for ethical AI systems, Wallach and Allen's *Moral Machines* (2009) devotes at least eighteen full pages to discussions of Asimov's Laws. In fact, the first pages of the introduction to the book list the laws and center them as a valid question to be further explored in the book. It's clear the authors recognize the laws are flawed and the product of fiction, but they still reach for them as a kind of cultural touchstone (and we should always ask whose culture any particular fiction works as a touchstone for). Before launching into a five-page discussion of the laws at one point, they say, "No discussion of top-down

morality for robots can ignore Asimov's Three Laws…" (91). And while they open their discussion with an acknowledgement that the laws are fiction, by the end of the book they are once again taking seriously the possibility that there could be robots programed with these laws as their ethical system (195).

A recent anthology designed to use science fiction to explore various philosophical topics devotes an entire chapter to Asimov's Laws and how they interact with machine metaethics, and one might even think it would be weird had they been excluded from the book (Anderson 2016). Indeed, in the now-classic 2012 anthology *Robot Ethics*, there are 13 separate mentions of Asimov's laws across nine different chapters (Lin – Abney – Bekey). In the updated version of the book in 2017, there are significantly fewer mentions of them, hopefully indicating that folks have begun to realize that they aren't a useful starting point, even if there is a lot to be said for reaching toward fiction to help frame our approaches to AI ethics (Lin – Jenkins – Abney). And again, while some of these references mention Asimov's Laws merely to dismiss them, they're mentioned often. The point here, of course, is not to attack Asimov or his laws, but that the way we think about AI ethics relates to the stories we tell (and the stories we read and hear), and the fact that Stanislaw Lem's work is rarely mentioned here means we've ignored or overlooked a rich depository of stories that would have helped us frame these questions much more successfully had we started there instead.

As far back as 2010, when they were drafted by some of the most well-known robot ethics researchers, one influential set of principles of robot regulation begins by listing Asimov's Laws, and then moving them aside as being not realistic or useful (Boden et al. 2017). Even people who work closely with robots and know that such laws cannot be consistently implemented feel the need to mention these laws just to move them aside. Indeed, it seems almost like people writing about robot ethics must address Asimov's laws to be taken seriously, while some roboticists do still insist on trying to implement them in actual machines (Torreson 2018; Van Dang et al. 2018; McGrath – Gupta 2018). In serious scholarship, from *Science* to *Nature*, Asimov's Laws make an appearance and guide real thinking in robot ethics, and have for some time (Deng 2015, Sawyer 2007). But it's more or less unheard of for serious robot or AI ethicists to cite Lem's fiction as a starting point. And that's a mistake. And it isn't just naivete that drives scholars to cite Asimov, it's the need for stories to frame our understandings, something Lem understood very well. And full disclosure: I, too, am guilty of citing Asimov to dismiss his laws in the context of robot ethics (Dietrich et al. 2021a, Dietrich et al. 2021b).

And it isn't just serious scholarly research that reaches for Asimov's completely untenable laws. Popular media does, too. In an article in Forbes magazine in 2018, one tech ethicist asks "Would Deviant Sex Robots Violate Asimov's Laws of Robotics?" (Lin 2018). In a 2017 article meant to be a bridge from academia to public outreach, a computer science professor opines that "After 75 Years, Asimov's Laws of Robotics Need Updating" (Anderson 2017). Writing a report for the company GE in 2016, another computer science professor writes "We Need Ethical Robots. Asimov's Laws are a Good Way to Start" (Kuipers 2016). When these kinds of fictional devices are what guides not just public opinion, but expertise as well, we end up with a great deal of confusion about what sorts of things robots can or should do, and a good deal of wasted time and energy in trying to work with or overcome ethical laws like this.

One reason for the constant invocation of Asimov's Laws in AI/robot ethics is that ethics (and AI ethics in particular) is a mess (Dietrich et al. 2021a). Professional ethicists generally recognize that the big historical systems we have to make sense of ethical theories can almost certainly not be held consistently. Try to be purely Kantian, purely consequentialist, and you'll eventually end up facing an action that seems obviously unethical on its face, but is ethical according to the system. Asimov's Laws, like those systems, is an attempt to make progress in spite of the impossibility of laying out such laws in advance. Ironically, of course, most people who have actually read a lot of Asimov's robot stories would immediately recognize that he in no way believed those to be good laws that would ethically bind autonomous robots. The stories are always framed around showing how a robot can be programmed with those laws and still break them (Asimov 1950; Asimov 1986). And yet, as we have already seen, in AI theory we still have many people who take them seriously as a starting point.

There are, of course, richer accounts of ethics than the simple rule-based systems we're all familiar with. There is a rich enactive literature around the ways that values are encoded for us in natural systems, and ways that ethics might emerge alongside things like consciousness (Torrance 2008; Thompson 2001). There are also rich accounts of things like moral imagination, growing out of pragmatist traditions (Johnson 1993). These embodied accounts of ethics require a much richer notion of a human than a simple rule-following system, and there's real deliberation involved in these ethics: not just calculation. Some of the most interesting work happening in AI ethics right now is coming out of an Aristotelian tradition, focused on the idea that we might be able to achieve ethical robots through something akin to artificial phronesis (Vallor 2016; Sullins 2016). These views are much less common in the fields of robot and AI ethics, but one could easily see Lem's robots testing the waters for similar kinds of views. In "Upside-down Evolution," for example, Lem seems exactly right when he writes, "Take, for example, the astonishing reversal – completely unforeseen – in the field of AI, which became a force to be reckoned with precisely because it did not become the machine embodiment of the human mind" (Lem 1986, 37–38). He tells us here about artificial nonintelligence as the path to building the kind of AI we actually want, understanding better than so many AI theorists even today that human intelligence might not be the benchmark we think it is. He writes: "Successive generations of information theorists and computer scientists had labored in vain to imitate the functions of the human brain in computers; stubbornly they ignored a mechanism a million times simpler than the brain, incredibly small, and remarkably reliable in its operation. Not artificial intelligence, but artificial instinct should have been simulated for programming at the outset" (51). This idea of artificial instinct giving us machine intelligence echoes back to Damasio's somatic markers and enactivism's claims that life and mind are a continuum. His skewering of GOFAI's high-level symbolic processing when he writes, "the majority of AI enthusiasts were still busy programming computers to carry on stupid conversations with not too bright people" (52) was a revelation that still smarts in many corners of the AI world, or would if more AI researchers knew Lem's work.

Lem has shown us that teaching a robot/AI ethics is exactly as easy or hard as teaching a human ethics, as long as we continue building AI with the goal of making them human-like. This seems like a small, even silly point. Of course, if we want human-like systems, they will be human-like! But far from silly, it seems most researchers in AI today just do not understand, or have willfully ignored, these implications in their work. Why are our NLP and facial recognition algorithms so racist? Because we, as humans, in this time and place, are racist. Why are our resume-analysis and

hiring algorithms sexist? Because we, as humans, in this time and place, are sexist (Flowers 2019, Birhane 2021). Lem was able to show us, with humor and grace, all of our human flaws, and we were only able to see them as robot follies. Hopefully, it's not too late to throw Asimov's laws out as a starting point and instead take up the nuance and humanity of Lem's robots instead.

Acknowledgements

The author thanks David Gunkel and Marcin Wichary, for comments and conversation about the difficulties of reading and analyzing Lem in translation.

Bibliography

Abid, A. – Farooqi, M. – Zou, J. (2021): Large language models associate Muslims with violence, *Nature Machine Intelligence* 3, 461–463, available at: < https://doi.org/10.1038/s42256-021-00359-2 >.

Anderson, S. L. (2016): Asimov's 'Three Laws of Robotics' and machine metaethics, in Schneider, S. (ed.) *Science Fiction and Philosophy: From Time Travel to Superintelligence*, Wiley Blackwell, 290–307.

Anderson, M. R. (2017): After 75 years, Isaac Asimov's three laws of robotics need updating, *The Conversation* [online] 2017-03-17, [accessed 2021-09-27] available at: < https://theconversation.com/after-75-years-isaac-asimovs-three-laws-of-robotics-need-updating-74501 >.

Asimov, I. (1950): *I, Robot*, Gnome Press.

Asimov, I. (1986): *Robot Dreams*, Ace Books.

BBC News [online] (2021): Margaret Mitchell: Google fires AI ethics founder, 2021-02-20 [accessed 2021-09-27] available at: < https://www.bbc.com/news/technology-56135817 >.

BBC News [online] (2020): Timnit Gebru: Google staff rally behind fired AI researcher, 2020-12-20, [accessed 2021-09-27], available at: < https://www.bbc.com/news/technology-55187611 >.

Bender, E. et al. (2021): On the dangers of stochastic parrots: Can language models be too big?, *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021-03-03, 610–623, available at: < https://doi.org/10.1145/3442188.3445922 >.

Birhane, A. – van Dijk, J. (2020): Robot Rights? Let's Talk About Human Welfare Instead, *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 2020-02-7, 207–213, available at: < https://doi.org/10.1145/3375627.3375855 >.

Birhane, A. et al. (2021): The Values Encoded in Machine Learning Research, *Cornell University arXiv:2106.15590 [cs.LG]* [online] available at: < https://arxiv.org/abs/2106.15590 >.

Birhane, A. (2021): The impossibility of automating ambiguity, *Artificial Life* 27, 44–61, available at: < https://doi.org/10.1162/artl_a_00336 >.

Boden, M. et al. (2017): Principles of robotics: regulating robots in the real world, *Connection Science*, 29(2), 124–129, available at: < https://doi.org/10.1080/09540091.2016.1271400 >.

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

Damasio, A. (1994): *Descartes' Error: Emotion, Reason, and the Human Brain*, Quill/Harper Collins.

Deng, B. (2015): Machine ethics: The robot's dilemma, *Nature* 523, 24–26, available at: < https://doi.org/10.1038/523024a >.

Dennett, D. (1984): Can machines think?, in Shafto, M. G. (ed.) *How We Know*, Harper and Row.
Dietrich, E. et al. (2021a): *Great Philosophical Objections to Artificial Intelligence: The History and Legacy of the AI Wars*, Bloomsbury Academic.

Dietrich, E. et al. (2021b): The AI Wars, 1950–2000, and their consequences, *Journal of Artificial Intelligence and Consciousness*, available at: < https://doi.org/10.1142/S2705078521300012 >.

D'Onfro, J. (2019): Google scraps its AI ethics board less than two weeks after launch in the wake of employee protest, *Forbes* 2019-04-4, [accessed 2021-09-27], available at: < https://www.forbes.com/sites/jilliandonfro/2019/04/04/google-cancels-its-ai-ethics-board-less-than-two-weeks-after-launch-in-the-wake-of-employee-protest/?sh=15b1bdab6e28 >.

Flowers, J. (2019): Rethinking algorithmic bias through phenomenology and pragmatism, *Computer Ethics – Philosophical Enquiry (CEPE) Proceedings*, available at:
< https://doi.org/10.25884/mh5z-fb89 >.

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

Kandel, M. (1992): Introduction, in Lem, S. *Mortal Engines*, Harvest/Harcourt Brace Jovanovich, i–vii.

Kuipers, B. (2016): We need ethical robots. Asimov's laws are a good place to start, *General Electric* [online] 2016-06-27, [accessed 2021-09-27], available at: < https://www.ge.com/news/reports/beyond-asimov-how-to-plan-for-ethical-robots >.

Lem, S. (1974) *The Cyberiad*, translated by Kandel, M., Avon Books.

Lem, S. (1977/1992) *Mortal Engines*, translated by Kandel, M., The Seabury Press/Harvest-Harcourt Brace Jovanovich.

Lem, S. (1986): *One Human Minute*, translated by Leach, C., Harvest-Harcourt Brace & Co.

Lin, P. (2018): Would deviant sex robots violate Asimov's law of robotics? *Forbes* [online] 2018-10-15, [accessed 2021-09-27], available at: < https://www.forbes.com/sites/patrick-lin/2018/10/15/would-deviant-sex-robots-violate-asimovs-law-of-robotics/?sh=3b7c612e6b42 >.

Lin, P. – Abney, K. – Bekey, G. (eds.) (2012): *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press.

Lin, P. – Jenkins, R. – Abney, K. (eds.) (2017): *Robot Ethics 2.0: from autonomous cars to artificial intelligence*, Oxford University Press.

Lohr, S. (2021): What ever happened to IBM's Watson?, *The New York Times* [online] 2021-07-16, [accessed 2021-09-27], available at: < https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html >.

McGrath, J. – Gupta, A. (2018): Writing a Moral Code: Algorithms for Ethical Reasoning by Humans and Machines, *Religions* 9(8), 240, available at: < https://doi.org/10.3390/rel9080240 >.

Newell, A. – Simon, H. (1976): Computer science as empirical inquiry: Symbols and search, *Communications of the ACM* 9(3) 113–126.

Olson, P. (2021): Much 'artificial intelligence' is still people behind a screen, in *Bloomberg* [online] 2021-10-13, [accessed 2021-10-13], available at: < https://www.bloomberg.com/opinion/articles/2021-10-13/how-good-is-ai-much-artificial-intelligence-is-still-people-behind-a-screen >.

Sawyer, R. (2007): Robot Ethics, *Science* 318(5853), 1037, available at: < https://doi.org/10.1126/science.1151606 >.

Shamir, L. (2020): A case against the STEM rush, *Inside Higher ED* [online] 2020-02-03, [accessed 2021-09-29], available at: < https://www.insidehighered.com/views/2020/02/03/computer-scientist-urges-more-support-humanities-opinion >.

Sullins, J.P. (2015): Applied professional ethics for the reluctant roboticist, *Proceedings of the Emerging Policy and Ethics of Human-Robot Interaction Workshop at HRI* [online], available at: < http://openroboethics.org/hri15/wp-content/uploads/2015/02/Sullins_Applied-Ethics-for-Reluctant-Robticists.pdf >.

Sullins, J. P. (2016): Artificial Phronesis and the social robot, in Seibt, J. – Nørskov, M. – Schack Anderson, S. (eds.) *What Social Robots Can and Should Do*, IOS Press, 37–39.

Thompson, E. (2001): Empathy and consciousness, *Journal of Consciousness Studies* 8(5), 1–32.

Torrance, S. (2008): Ethics and consciousness in artificial agents, *AI & Society* 22, 495–521.

Torreson, J. (2018): A Review of future and ethical perspectives of robotics and AI, *Frontiers in Robotics and AI* [online] 2018-1-15, [accessed 2021-09-28], available at < https://www.frontiersin.org/articles/10.3389/frobt.2017.00075/full >.

Turing, A. (1950): Computing machinery and intelligence, *Mind* (LIX)236, 433–460.

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

Van Dang, C. et al. (2018): Application of modified Asimov's laws to the agent of home service robot using state, operator, and result (Soar). *International Journal of Advanced Robotic Systems* 15(3), available at: < https://doi.org/10.1177/1729881418780822 >.

Wallach, W. – Allen, C. (2009): *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press.

Weizenbaum, J. (1966): ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine, *Communications of the ACM*9, 36–35, available at: < https://doi.org/10.1145/365153.365168 >.

Wittgenstein, L. (1953): *Philosophical Investigations*, Macmillan.

Zakaria, F. (2015): Why America's obsession with STEM education is dangerous, *Washington Post* [online] 2015-03-26, [accessed 2021-09-28], available at: < https://www.washingtonpost.com/opinions/why-stem-wont-make-us-successful/2015/03/26/5f4604f2-d2a5-11e4-ab77-9646eea6a4c7_story.html >.