Human Rights, Artificial Intelligence and Heideggerian Technoskepticism: The Long (Worrisome?) View

Faculty Research Working Paper Series

Mathias Risse
Harvard Kennedy School

February 2019

RWP19-010

CARR CENTER FOR HUMAN RIGHTS POLICY

Human Rights, Artificial Intelligence and Heideggerian Technoskepticism: The Long (Worrisome?) View

Mathias Risse is Lucius N. Littauer Professor of Philosophy and Public Administration and Director of the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University. His work primarily addresses questions of global justice ranging from human rights, inequality, taxation, trade and immigration to climate change, obligations to future generations and the future of technology. In addition to the Harvard Kennedy School, he teaches in Harvard College and the Harvard Extension School, and he is affiliated with the Harvard philosophy department. Risse is the author of *On Global Justice* and *Global Political Philosophy*, as well as two other forthcoming books.

HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

DISCUSSION PAPER
FEBRUARY 2019

Human Rights, Artificial Intelligence and Heideggerian Technoskepticism: The Long (Worrisome?) View

Mathias Risse, Harvard University

February 11, 2019

1. Introduction:

By way of diagnosing the current cultural and scientific state of affairs, Nietzsche famously spoke of the "death of God."[1] The key to that ominous phrase is that anything that can die must once have lived. God is said to have lived at some point, but then perished. Belief in a god who represented a metaphysical reality beyond human perception was warranted for some period but no longer is. What according to Nietzsche brought about that change is our inquisitiveness about what the world is *really* like, triggered by the very stipulation of such a god. Ultimately that god himself became incredible. Inquiry in particular erodes sources of moral certainty, making us scramble to identify new ones.[2] Philosophers have endeavored to do just that. There is a wealth of proposals to make sense of our moral lives that do not depend on theology. But the current intellectual climate is such that the burden is distinctly on those who wish to theorize notions such as justice or also human rights. They must persuade those who believe the death-of-God metaphor reaches much beyond theology and altogether undermines our ability to make moral demands against each other.[3]

---

[1] I presented versions of this material at the conference on Human Rights, Ethics and Artificial Intelligence at the Harvard Kennedy School (jointly sponsored by the Carr Center for Human Rights Policy, the Edmond J. Safra Center for Ethics and the Berkman Klein Center for Internet and Society at Harvard) and the Harvard human rights colloquium, both in December 2018. I'm grateful to both audiences for discussion.

[2] Nietzsche, *The Gay Science*. See especially sections 125 and 343.

[3] The view that philosophy is incapable of establishing any grounding for either duties or rights is defended vigorously in Bittner, *Bürger sein*. According to Bittner, we live in a demand-free world. I do not agree with that view and have tried to show how to establish such a grounding, see Risse, *On Global Justice*. But whatever

A broadened understanding of the death-of-god claim is useful to characterize our intellectual situation when it comes to long-term challenges around artificial intelligence (AI). On the one hand, scientific inquiry has brought AI on our radar. On the other hand, it is precisely because we have made so much progress that we also increasingly grasp the limits of our understanding elsewhere. Lack of agreement on moral foundations is a case in point. Another is our comprehension of something very basic, *consciousness.* Is the mind just the brain, or is there more to it? If so what? If the mind just is the brain, does this mean the mind is an entirely physiological phenomenon?[4] On any of these possibilities, could machines have minds? Much as moral philosophy continues its disagreements about foundations of morality so philosophy of mind is divided about our mental self-understanding. The broader understanding of the death-of-god challenge is, in a nutshell, that knowledge enables but also throws into doubt. While we increasingly know how to build smart machines, we do not know what kind of morality to code into them, nor what kind they will gravitate towards once their own learning kicks in. Nor can we be sure if machines can end up having minds.

Barely does a day go by without reporting on this or that innovation classified as AI. Such reporting normally concerns *special* AI, smart machines with some capacity to learn on their own. For the taste of many engineers there is too much hype about *general* AI, which approximates or surpasses natural intelligence across a broad range of contexts. They

---

else is true, Bittner's book makes painfully clear that not only politically but also and especially philosophically the burden is on those who wish argue to for moral duties or rights of any sort.

[4] For consciousness see Blackmore and Troscianko, *Consciousness*; Blackmore, *Conversations on Consciousness*. For philosophy of mind see Jaworski, *Philosophy of Mind*; Heil, *Philosophy of Mind*; Braddon-Mitchell and Jackson, *Philosophy of Mind and Cognition*; Carter, *Minds and Computers*.

remind us that we (that is, *they*) are nowhere near developing something like that.[5] But "nowhere near" means "in terms of current technology." Innovation may quickly revolutionize how things are done. How much time will pass before major breakthroughs occur is a question that has no good answer. We must reflect on the impact of technology even though we have no clear idea when it will arrive, if ever.

My concern is with the impact of AI on human rights. Section 2 introduces a time-axis, distinguishing short, medium and long-term. The distinction between medium and long term – and those terms are our topic here -- might be more analytical than temporal but is nonetheless significant. The medium term is when new technologies will have thoroughly transformed society but machines themselves do not yet come up for moral consideration. The long term is the when such technologies will come up for such consideration. So we are then talking about a time where we would share our lives not only with animals but also with machines of various types that arguably would have features that render them both morally relevant and less acquiescent than animals to whatever abuse humans inflict upon them.

Section 3 identifies two presumptions about ethics-and-AI we should make only with appropriate qualifications. These presumptions are that (a) for the time being investigating the impact of AI, especially in the human-rights domain, is a matter of investigating impact of certain *tools*, and that (b) the crucial danger is that some such tools – the artificially intelligent ones – might eventually become like their creators and conceivably turn against them. We turn to Heidegger's influential philosophy of technology to argue these

[5] In 1950, Alan Turing thought we were only a few decades way from developing such general AI, see his seminal Turing, "Computing Machinery and Intelligence." A hallmark for such general AI is that it could play the "imitation game," and thus deceive human interlocutors who cannot see whether the partner in the conversation is a human or a machine.

presumptions require qualifications of a sort that should inform our discussion of AI. First of all, understanding technology merely as a set of tools would enormously underestimate how for each generation technology shapes the way we live and how we perceive fellow humans and other entities in the world. Technology has ways of enabling us in many ways while also impoverishing human life in crucial ways. But secondly, on the reassuring side, there is much more than intelligence to a proper understanding of human life. So at least it will be much harder to substitute us than it would be if that were merely a matter of building intelligence. But this reassurance also entails that "moral status" might be a more complex matter in the future than it is now. Human rights are meant to protect all and only humans; everything else has some other moral status. In the long term this might be too simple a picture. Heidegger's philosophy allows us to make these points in a unified manner and thus sets the stage for our investigation of the medium and long term.

Section 4 turns to the medium term. As far as human rights are concerned, one major challenge is how human rights will prevail in an era that quite possibly is shaped by an enormous increase in economic inequality. In admittedly bleak but not implausible scenarios technological innovation not only exacerbates inequality but renders large parts of the population superfluous. With ownership of data left unregulated wealth and power of tech companies might increase enormously, with no guarantees it will be used for social benefit. Currently the human-rights movement is rather unprepared to deal with these challenges. What is needed is greater focus on social justice/distributive justice, both domestically and globally, to make sure societies do not fall apart. As far as this medium term is concerned, we must also heed an admonition we draw from Heidegger that technology must be used to enhance the distinctively human life rather than impoverishing it.

Section 5 turns to the long term, a period when machines themselves come up for moral consideration. Even more than for the medium term we are woefully underprepared for this period, in a way the death-of-God theme makes clear. We do not have settled views on when artificial minds would be sufficiently similar to a human one be eligible for the similar moral status. What we can say, however, is that we must be prepared to deal with more types of moral status than we currently do and that quite plausibly some machines will have some type of moral status, which may or may not fall short of the moral status of human beings (a point also emerging from the Heidegger discussion). Machines may have to be integrated into human social and political lives. We will be better able to do so if we integrate humans more properly in the medium run, through greater focus on distributive justice, both domestically and globally, as discussed in our exploration of the medium run.

2. Short, Medium and Long Term

A time axis helps to differentiate among various ways for technology to have an impact on human life and human rights specifically: short, medium and long term. The short term is what already goes on, and is subject to scientific investigation or legal analysis. The Universal Declaration reflects experiences of the industrial age: there is a state that at least potentially provides services and has capacities to decide how to do so, a society with a public sphere of sorts where opinions are spread and debated, a workforce that is partially organized and a legal apparatus sufficiently fine-grained to also generate many possibilities for abuse. Changes in technology inevitably affect how rights designed for such a society can

be exercised, partly because they affect what states and other powerful entities can do to interfere with them. Inquiries about the short-term ask about such change.[6]

The medium term is the period when new technologies will have thoroughly transformed society but machines themselves do not yet come up for moral consideration. Precisely what such a period will look is hard to envisage since much technological innovation giving rise to yet more innovation will have occurred by then. Plausibly the nature of work will change. Algorithms embedded into appropriately designed devices with access to enormous amounts of data will be integrated into many lines of work. They will eliminate or radically change many jobs where they currently merely assist humans. In the medical domain there will be progress in prevention, diagnosis and treatment. Gene-editing is already on the horizon. Prenatal selection may noticeably affect the composition of the population, over generations plausibly making the wealthy on average distinctly more attractive by all normal standards than then poor. Warfare too may change through drones and automated weapons-systems.

This is a period beyond the ken of present social-scientific research. Historical learning is also only of limited value here. Innovation will lead to more innovation, leading to yet more innovation. In between people will change alongside technology. So far technological innovation has on balance always created more jobs than it has annihilated. And quite possibly this will be so again. Technological innovation creates new challenges and generates needs for new products and activities. It also frees up resources so people can be deployed in hitherto neglected domains. In benign scenarios new technologies would

[6] For a sophisticated report (prepared for the Canadian government) on the impact of AI on human rights, see Raso et al., "Artificial Intelligence & Human Rights: Opportunities & Risks."

generate widely dispersed benefits, allowing people a leisure to develop their personalities or simply follow their interests that historically only a selected few could enjoy.

But there is no guarantee that all this will be so, and some reason to think it will not. Technology might also create a situation where many people are no longer needed in society at all, and where the wealthy refuse to share what they can accumulate without cooperation of others and instead withdraw to gated communities where food production, transportation and upkeep is left to machines, liquidating the jobs previously associated with such work. When wealth was determined by land ownership, those who did not own land were needed as tenants, and the possibility of renting to tenants was the point of land-ownership. When wealth was determined by ownership of factories, those who did not own factories were needed to operate machines and buy products thus created. There was a point to machine-ownership only because there were such people. But once ownership of data matters most for wealth-creation and sophisticated algorithms help to get things done, many people may no longer be needed at all. And once they are no longer needed for production or consumption, their claims to any kind of political consideration as members of society might go unheeded. As of a 2018, in a world riddled with human rights challenges, some people buy adult albino dragonfish for $70,000, billionaires dream of going to the moon as tourists, and tax evasion among the wealthy is rampant around the world.[7] This is not a world where anybody should be confident that solidarity or moral considerations would carry much weight once the lives of the rich become more detached.

[7] On the dragonfish, see *Economist*, Sept 15, 2018, p 40. On tax evasion see Dietsch, *Catching Capital*.

The long term is when machines have become so sophisticated that we must wonder if they come up for moral consideration on their own terms. One possibility is that there will be cyborgs (cybernetic organisms), organic beings with integrated artificial parts. Another possibility is that increasingly sophisticated machines will ever more look and act like humans, creating an android population alongside (genetically enhanced) humans. Perhaps digitalized brains could be uploaded on computers and thus be preserved. Or there could be sophisticated machines that are nothing like humans. At some point such entities might surpass humans in just about all mental and many other capacities, and will then also be able to design yet other entities that surpass *them*. This moment is known as the *singularity*.[8] Such a moment would have a dramatic impact but at a much earlier stage we might already end up with machines that must morally be considered on their own terms, and be taken seriously as at least co-inhabitants of sorts of the same social space as humans. When we come to a point where the main difference between "us" and "them" is that we are made from carbon and they are not, we would indeed have to take the idea very seriously that they deserve a more serious kind of consideration than they normally do or should receive now.

These questions are both moral and political in nature. Humans have managed to keep other animals in control by domesticating them to enjoy their company or harvest their bodies, putting them on display on zoos or corralling them into wildlife preserves. But smart machines might not forever be that docile, and develop capacities to rebel again their human creators.[9]

---

[8] For the view that the singularity is, well, near, see Kurzweil, *The Singularity Is Near*. For philosophical discussion, see Chalmers, "The Singularity: A Philosophical Analysis."

[9] For intriguing speculations about what kind of living arrangements there might be between humans and machines, see Tegmark, *Life 3.0*, chapter 5.

3. Borrowing from Heidegger on Technology

By way of preparing us for our investigation of the medium and long-term we now identify two presumptions we should only make without substantial qualifications. These presumptions are that (a) for the time being investigating the impact of AI, especially in the human-rights domain, is a matter of investigating the impact of certain *tools*, and that (b) the crucial danger is that some of these tools – the artificially intelligent ones - in the long run might become too much like their creators and conceivably turn against them. We turn to Heidegger's influential philosophy of technology to argue that these presumptions require substantial qualifications to assume a plausible form.[10]

Heidegger is off-putting for two reasons. For one thing, his writings, especially his 1927 main work *Being and Time*, are forbidding, littered with a technical vocabulary taken from everyday German but then used terminologically and hard to translate into other language.[11] But here one should be patient (and get assistance from a good introductory book) since Heidegger's purpose is to reflect on just what it is for us *to be in the world*. This task is not what our common vocabulary is normally used for. The other reason is his involvement with the Nazis and his genuine sympathies for some of their guiding themes,

[10] For this section I have benefited from Young, *Heidegger's Later Philosophy*. See also Richardson, *Heidegger*; Wrathall and Critchley, *How to Read Heidegger*. For Heidegger on technology see also Borgmann, *Technology and the Character of Contemporary Life*; Dreyfus, *What Computers Can't Do*; Dreyfus, *What Computers Still Can't Do*. On Heidegger and the philosophy mind see also Olafson, *Heidegger and the Philosophy of Mind*. For the connection to cognitive science, see Kiverstein and Wheeler, *Heidegger and Cognitive Science*.

[11] Heidegger, *Being and Time*.

especially Anti-Semitism.[12] But though these are very real concerns, they must not deceive us about the pathbreaking importance of Heidegger's thought, especially for a proper understanding of human life in a technological era. This point is only reinforced by the fact that Heidegger's premier American propagator, Hubert Dreyfus (a Berkeley philosopher who passed away in 2017) was an early and influential critic of AI. Dreyfus articulated his criticisms from a distinctly Heideggerian standpoint in ways that over time increasingly many people have seen as valid.[13]

Heidegger is a major representative of a philosophical movement known as *phenomenology.* Phenomenologists share a fascination with the complexities and enigmatic nature of human experiences and resist their oversimplifications in philosophy and the sciences. Heidegger himself resisted much traditional philosophy by focusing on the concrete living-experiences of human beings in their respectively thick contexts. An informative contrast is with Descartes, who understood humans essentially in terms of their thinking abilities. But what is distinctive about the human manner of being in the world, according to Heidegger, is not that we *reflect* about the world, but that we *care about* other things whose meaning is established through use patterns in a shared life world.[14]

Heidegger uses the term *Dasein, there-being,* to capture our inherently social existence as being who always already possess a pre-theoretical grasp of the a priori structures that make possible particular ways of being in the world. The possibilities we

[12] For recent discussions of these matters, see Espinet et al., *Heideggers "Schwarze Hefte" im Kontext*; Wolin, *The Politics of Being*; Mitchell and Trawny, *Heidegger's Black Notebooks*.

[13] Dreyfus, *What Computers Can't Do*; Dreyfus, *What Computers Still Can't Do*; Dreyfus, *On the Internet*.

[14] For the importance of Descartes for the contemporary philosophy of mind, see Heil, *Philosophy of Mind*, chapters 2-3.

understand are those whose meaning we share with everyone else. To use a simple example, the hammer's being in our human life-world is its readiness-to-hand, its role in our world. Its "true" weight as understood in terms of its place in our world is its being too light or too heavy to use effectively, not a neutral one or two pounds. Its true place in that same sense is the fact that it is too near or too far away to use well, not a point on a geometric grid. Much like our ways of integrating hammers into our life-worlds, more generally our use-patterns, movements, choices and thoughts are thus usually merely instances of what "we" are and "they" do or choose or believe. "Authenticity" then becomes important, which is a truthful relationship to our "thrownness" into a world with which we are always already concerned.

Thinking itself is always already thinking *about* things discovered in everyday practical engagements. Our understanding of the world is not reducible to a brain state, but instead constituted by a whole body being integrated into a world of objects and relations interconnected in complex ways. This approach inspired influential criticism of scientific approaches that did reduce our being in the world to a brain state. For instance, French philosopher Maurice Merleau-Ponty objected to mainstream psychology, insisting that body skills are *constitutive of* and not merely somewhat *directed by* intelligence.[15] Based on both Heidegger's general outlook and Merleau-Ponty's elaboration, Dreyfus has argued that disembodied symbol manipulation performed by computers cannot realistically hope to reproduce the flexible and nuanced competent behavior human beings (as well as other animals) display routinely.[16]

[15] For an introduction, see Carman, *Merleau-Ponty*.

[16] John Haugeland rebutted "Good Old Fashioned AI" along these lines, and recently Alva Noë, in *Out of Heads*, suggested that rather than being something that happens inside us, consciousness itself is something we do. Haugeland, *Artificial Intelligence*; Noë, *Out of Our Heads*.

So one major theme in Heidegger is the embeddedness of *Dasein*, which generates a certain attitude towards AI, namely, a confidence that replacing humans with machines is not as straightforward as those may have thought who, in Cartesian spirit, saw humanity's essence in abilities to think, which was then increasingly seen as something that, one way or another, could be reduced to computation. With a that reduction established, approximating the human mind becomes an exercise of developing a suitable algorithm to do the computations. But what is implied by Heidegger's approach is that for machines to become smarter and smarter will not mean they ipso facto become increasingly human-like. It takes much more than intelligence to do so.

Another theme in Heidegger is how the ways in which we are embedded into our social world open up certain possibilities while closing off others. It is because of certain background assumptions and understandings that we know what is under discussion. This is true both individually (each person is embedded into a life-world that provides an understanding of relations and objects around her) and collectively, where how societies share particular ways of relating to the world excludes other ways of relating. Heidegger calls such background understandings "horizons of disclosure." His understanding of truth also enters here, which amounts to bringing entities out of hiddenness so we can deal with or make correct statements about them. He also talks about ways in which the world is "unconcealed" or "revealed," which entails there are ways in which the world remains concealed. One danger is that perspectives become narrow in such a way that their horizontal character is forgotten, a phenomenon Julian Young called the "absolutization" of

one horizon.[17] In the present age it is technology – or more specifically the ways in which technology is integrated into our lives – that determines our horizon of disclosure.

In his 1953 *The Question Concerning Technology*, Heidegger takes on the ordinary understanding of technology as means to ends and as a product of human activity.[18] Technology, on that view, is a set of tools we control, a vast array of instruments, artefacts and devices. But while this understanding is sufficient for many purposes it does not do justice to technology's significance in human life. Technology also makes things show up as mattering in some way or other and thereby is a mode of revealing. The mode of revealing characteristic of modern technology understands everything around us to be no more than what Heidegger calls *standing-reserve*, resources to be exploited as means. This includes all components of the natural world, even humans, who are deployed for other people's purposes in all sorts of way. In 1966 Heidegger predicted that "some day factories will be built for the artificial breeding of human material (…) according to plan and need."[19] There do not seem to be such factories, but there is a depressing amount of trafficking of humans who are reduced to a kind of readiness-to-hand, much like a hammer.

Heidegger famously uses the example of a hydroelectric plant on the Rhine that converts the river into a mere supplier of water power. As a contrast he offers an old wooden bridge that spanned the river for hundreds of years, which reveals it as a natural environment and permits a kind of poetic habitation where natural phenomena are revealed

[17] Young, *Heidegger's Later Philosophy*, 29.

[18] Heidegger, *The Question Concerning Technology, and Other Essays*.

[19] Quoted in Young, *Heidegger's Later Philosophy*, 46.

to us as objects of respect or wonder. Technological revealing is not a peripheral matter but defines our modern way of living, at least in the West. The problem is that technological revealing extinguishes any sense of awe and wonder, a loss to which, Heidegger worries, we are essentially indifferent. He uses the term *Gestell*, often translated as *enframing*, to capture the relevance of technology in our lives. The *Gestell* is a horizon of disclosure according to which everything registers only as a resource. *Gestell* drives out our ability to see the what-ness of things, their in-itself-ness, depriving us of the ability to stand in a caring (care-ful), rather than violent, relation to things. In one striking formulation Heidegger pointed out that the modern world is revealing itself as a "gigantic petrol station."[20] Trees become sources of cellulose, rivers suppliers of power, even unspoiled areas of nature are reduced to resources ripe for exploitation by tourism and people become sexual resources. *Gestell* makes violation essential while making us oblivious to what is going on.

For any of this to be illuminating philosophical thought rather than retrograde agrarian nostalgia we must not read it as Luddite rejection of technology. The way forward, according to Heidegger, is to inhabit technology differently. Technology needs to be there for us to enjoy and use, but is not our only or fundamental way of encountering entities. Heidegger has his own views on what this would mean (known as the ethics of dwelling) but we can leave him at this stage. The point is that technology turns us into certain types of humans that can relate to the world only in very impoverished ways. Everything is interconnected and exchangeable, efficiency and optimization setting the stage. Efficiency demands standardization and repetition. Technology safes us from having to develop skills,

[20] Quoted in Young, 50.

and also makes us become the kind of people satisfied that way. Instead, one way or another, we need a receptivity to, and a desire to care about, things as they are independently of us.

So, in sum, Heidegger is of interest to us because he aimed to refocus what philosophy should be all about, and how we should understand being in the world, just about when technological breakthroughs set us on the path to an entirely new type of technology that grew out of precisely that approach to philosophy that made the mind central that Heidegger tries to leave behind. Technology is not well-understood merely as a set of tools but as setting the general limits on how humans see themselves in the world, possibly to the exclusion of ways of being in the world that do not sit well with the dominant role of technology. We must pay heed to making sure technology will not end up making humans dumber, less creative or excessively over-specialized.

But Heidegger also offers the reassuring insight that there is much more than intelligence to a proper understanding of human life. At the very least it will be much harder to substitute us than it would be if it were merely a matter of building intelligence. However, this reassurance also entails some complexities, to wit, that "moral status" might be a much more complex matter in the future than it is now. After all, intelligent and otherwise capable machines might come up for moral considerations well short of having arrived at a stage where they are somehow equivalent to humans. But not being equivalent to humans at some stage can no longer mean the entity in question does not come up for moral consideration. Human rights are meant to protect all and only humans; everything else has some other moral status. In the long term this might well be too simple a picture. Heidegger's philosophy allows us to make these points in a unified manner and thus sets the stage for our investigation of the medium and long term.

4. Human Rights and AI in the Medium Term

Many philosophers and activists prefer to treat rights separately from distributive justice. Both pragmatically and philosophically speaking, rights, especially human rights, seems to be beyond contestation in ways in which distributive justice never could be. But such an approach fails to make rights, especially human rights, a comprehensive or even stable ideal worth aspiring to. The reason, in short, is that increasing technological advancement would not only generate yet more staggering inequality, but might render large parts of the population redundant. Considerations of social and global distributive justice would then lose their grip, and alongside them, human rights might as well. The basis on which one could ask for distributive justice or human rights, as well as the motivational structures for people to want to deliver on them, might disappear or be reduced drastically. If we want to prevent such scary scenarios, rights now need to be supplemented with broader considerations of social justice domestically and global distributive justice internationally.

For vividness consider the 2013 American science fiction movie *Elysium.* Set in 2154, Earth is overpopulated and polluted. Most people live on the edge of starvation, with little access to technology and medical attention. The wealthy live on a gigantic space habitat in Earth's orbit called Elysium. Elysium is technologically advanced, with devices that cure diseases and reverse aging. Needless to say, there is much hostility between Earth and Elysium. The way to avoid such scenarios would be to strengthen social and global distributive justice now, to nip in the bud any kind of development that would lead to such a scenario. That is the thought this section develops.

An early statement of how rights constitute an insufficient ideal for human emancipation appears in Karl Marx's 1843 treatise *On the Jewish Question*. Ostensibly Marx is assessing his contemporary Bruno Bauer's reflection on attempts by Jews to achieve political emancipation in Prussia. True political emancipation, according to Bauer, requires abolition of religion, rather than protection of particular religious orientations. In response Marx argues that Jews could achieve political emancipation in terms of rights just fine without renouncing religion. But he questions the potential of rights to create genuinely human emancipation. The "so-called rights of man," Marx points out, are "rights of the member of civil society, i.e. egoistic man, man separated from the other men and the community."[21] Rights keep people isolated from each other, protecting them only in their pursuit of their private interests, especially preservation of private property.

For Marx, human emancipation will only be complete when "as an individual man in his empirical life, in his individual work and individual relationships becomes a species-being."[22] Human species-being, suffice it to say, is about a communally richer experience than a society focused on political emancipation in terms of rights would make possible. Ours is an impoverished social world if that world is structured by rights, which is a point about the relationship of human beings among each other within society. This dovetails with but is also importantly different from Heidegger's concern about our social world overall being able to reveal only a limited range of possibilities that there are in the world.

[21] McLellan, *Karl Marx: Selected Writings*, 60.

[22] McLellan, 64.

It is a notoriously difficult question to assess how important Marx thought considerations of distributive justice should be. Any kind of moral talk Marx normally relegated to the ideological *Überbau* of society, ideas that merely reflected class interests. Instead, he made his point in terms of human species-being. A century later, however, John Rawls echoed the basic gist of Marx's point, but he now did so in terms of distributive justice.[23] Rawls famously formulates two principles of justice. The first talks about civil and political rights, demanding each person have as broad a range of them as compatible with any other person having the same. But the second principle supplements the first by demanding fair equality of opportunity as well as a distribution of wealth and income in society in a manner that pays special attention to the least advantaged. Like Marx, Rawls insists on a richer ideal of communal life, beyond what rights can accomplish. But unlike Marx he did so explicitly in terms of social justice.

But this much has been about the *domestic* level. More recently, as far as the international level is concerned, historian Samuel Moyn has questioned the emancipatory value specifically of human rights.[24] Moyn argues that the human-rights movement succeeded while other political options were also on the table, including a UN framework much more concerned with economic integration and empowerment of developing countries. In other words, the human-rights movement prevailed to the detriment of broader social and economic justice. Human-rights activists sought remedies for destitution without challenging wealth, so without trying to achieve a more equitable and to that extent more

[23] Rawls, *A Theory of Justice*.

[24] Moyn, *Not Enough*.

humane world. My own approach to global justice is broadly supportive of Moyn's critique. My book *On Global Justice* embeds human rights into a larger structure of global distributive justice. My forthcoming *On Trade Justice: A Philosophical Plea for a New Global Deal* elaborates on these themes within the domain of trade.[25]

To see how all this matters for our current discussion, recall the debate about Thomas Piketty's *Capital in the 21st Century.*[26] Piketty's point is that the inner workings of capitalism are such that, in times of peace, those who own the means of production will benefit proportionately more from the economy than those who merely work there. Over time, from generation to generation, the result is going to be an enormous increase of wealth inequality, not due to differences in the wage structure but to inherited wealth. We did not observe this during the 20th century because the Great Depression and the two global wars created so much calamity and destruction that these tendencies remained hidden. More recently historian Walter Scheidel has added his analysis that, historically speaking, serious pollical efforts to reduce inequality were only ever successfully taken in times of great calamities such as economic collapse and outbreaks of epidemics.[27] These were times when the political momentum could be generated to bring about lasting economic change. So the experience of the 20th century is representative. Putting these perspectives together: there will either be more and more inequality, or not. But if not, it will be because of some such a calamity. Either way, this would not be a world many of us would happily inhabit.

---

[25] Risse, *On Global Justice*. Risse and Wollner, *On Trade Justice: A Philosophical Plea for a New Global Deal*.

[26] Piketty, *Capital in the Twenty-First Century*.

[27] Scheidel, *Great Leveler*.

Recall *Elysium*. Technology will likely exacerbate tendencies towards more inequality that are inherent in capitalism. Those who help create new technology or know how to turn it to their advantage have enormous earning potential. Since it normally requires much education to produce or benefit from new technology, inherited privilege will make it easier to end up with such work, and then earnings will enlarge inequality more. Increasingly, highly educated people get married to others of such standing, yet more enhancing these tendencies. In the Marxian picture increasing inequality, or that is, indigence on the lower end of the economic ladder, eventually triggers a revolution. But there could be a revolution only because the working class would remain within the capitalist economy and was needed there. There could be no revolution if the working case were no longer part of the economy. It is that possibility that technological innovation dangles before us, with *Elysium* offering an eerily extreme vision of a future that is post-revolutionary in a very different sense than what Marx had in mind.

Post-Marxist thinkers had to address the challenge that the revolution Marx predicted never happened. One way or another, the explanation they offered was that capitalism succeeded in coopting the working-class within the system. Capitalism offered them just enough, and the right kind of thing, for any revolutionary energies they ever might have had to evaporate. In the process capitalism obtained a vast sea of customers for products people had no deeper need to buy. Their eagerness to stuff their lives with these products that replaced any kind of revolutionary energy. But it might all be a rather different story in the medium term. Many tasks performed by low-income workers can be outsourced to machines. Economic structures might change in such a way that it is ownership of data that constitutes wealth. The source of wealth no longer is that people buy stuff. In that case the

underclass might no longer be needed to keep the economy going. They might live in Bantustans, as people of color did in Apartheid South Africa, and eventually there might be something like an Elysium. The wealthy transpire into the sky and the under-privileged fade from view and thus would be out of mind as well, undermining any psychological willingness and political capacity to accommodate them.

Social justice is fueled by claims people make because they are cooperators in a vast and intricate economic system, and because they comply with rules. If they do all that, they have good reason to demand shares in the political economy. But if they are not economically integrated, they no longer have any political or philosophical grounds to clamor for social justice. "Social justice" is a term that only became applicable under the conditions of dense interconnectedness that the Industrial Revolution made possible. But eventually advancing industrialization might undermine that very interconnectedness its earlier stages created and based on which social-justice claims made sense.

Societal fragmentation must also be seen at a global level. Human rights claims are thinner than claims of domestic justice. Social justice, that is, requires more than a set of rights. But the fearsome scenario for the future we just explored would also entail that a number of reasons that currently allow us to argue for human rights would disappear. If there is no interconnectedness within societies, there is no interconnectedness at the international let alone global level. That in turn would undermine any kind of argumentative support for human rights that comes from global interconnectedness. Needless to say, there would still be claims that draw on common humanity alone. But those would likely be motivationally impotent absence global interconnectedness. The remedy for this is to emphasize talk about social justice and global distributive justice now, whose realization

would strengthen and develop into the future the societal interconnectedness that renders social justice and global distributive justice applicable in the first place. Social justice and global distributive justice should be promoted by anybody concerned with human rights. Focusing on human rights exclusively is neither philosophically plausible nor strategically advisable as far as this medium term is concerned.[28]

5. Human Rights and AI in the Long Term

The hallmark of the long term is that machines themselves may well come up for moral consideration. My point here is that it is indeed plausible that machines will be moral agents of sorts in the future, and that this will require radical rethinking of our societies in ways that includes machines as participants of sorts. We will best prepare for that if we start taking social-justice considerations more seriously at this stage. If *humans* are taken seriously now, *machines* can be more smoothly integrated into society when the time comes.

[28] (1) One way of avoiding the kind of exclusion of large parts of the population that is carried to an extreme in the *Elysium* scenario is not to have a population that would generate the exclusionary processes to begin with. In other words, if humanity shrank radically and in the process focused on providing the kind of education to everybody that would make them successful participants in a high-tech economy, the problems we discussed here might not arise. Having a smaller overall population might be a sensible response also to climate change. But needless to say, the moral and political challenges involved in such a transition would be formidable. (2) We must pay special attention to the increasing power of tech companies. Leading companies in the AI sector are already more powerful than oil companies ever were, and this is just the beginning of their ascension. If the power of companies such as Alphabet, Apple, Facebook or Tesla is not harnessed for the public good, we might eventually live in a world dominated by companies, as depicted in Margaret Atwood's *Oryx and Crake* or David Foster Wallace's *Infinite Jest*. Companies must be integrated into the struggle for social justice, and need to see themselves as responsible for doing so. The response to this point cannot be that companies are private actors we should not think of as beholden to the public good in any way. After all, any type of company, especially the modern corporation, is delineated in its structures and capacities by a legal framework that is provided politically. Corporations do not exist in a state of nature or otherwise just on their own. It is legitimate and profoundly important to rein them in for public purposes. This must concern especially ownership of data, which, if left unregulated, would be a major engine for inequality in the future. But there is no good reason to let it come to that. For the moral role of corporations, see Risse and Wollner, *On Trade Justice: A Philosophical Plea for a New Global Deal*, Part III.

Let me illustrate what is at issue with another movie, the 2018 Netflix production *Extinction*. Set at an unknown time in the future, we witness what appears to be a brutal alien invasion to extinguish humanity. Only gradually do viewers understand that the invaders are humans who come to reclaim the planet from androids who had driven them into inter-planetary exile fifty years earlier and had self-implanted memories to create experiences of living in a human society. That earlier conflict occurred because androids had become human-like but were treated as tools – tools, however, that could increasingly *feel* humiliation. We see flashbacks at demonstrations where human mobs chant "you will not replace us." A key scene in the movie's present is when a human soldier cannot bring himself to kill a group of androids who think they are a family and display all relevant emotions.

We are woefully underprepared philosophically for anticipating super-intelligent machines, in at least two ways. To begin with, given that we do not have anything like agreement for how to think about morality we have no good way of predicting what a super-intelligence would do once its own learning kicks in and the issue is no longer about solving a value-alignment problem by creating a particular code. Responses to the death of god include positions ranging from Bittner's demand-free world via Rawlsian public reason that endorses the variety of ways in which humans have historically made sense of their moral lives to an ongoing adherence to moral realism. Or, for an additional position, recall Heidegger's view on ethics mentioned in passing. Heidegger turns his back on the abstractions of moral thought and wants to return to the concrete life worlds of contextually embedded caring beings, to an ethics of dwelling. Philosophers who are very intelligent by contemporary human standards are on all sides of this debate. At this stage it seems appropriate to try to draw lessons from the sheer fact that there is such a diversity of

approaches. But with some intellectual modesty one can really only say we are rather clueless as to how any kind of super-intelligence would process morality.[29] Maybe a super-intelligence will be super-confused from the variations among available approaches.[30]

We are also clueless when it comes to assessing what our relationship with machines should be once they assume new levels of sophistication. For in crucial ways we do not understand ourselves and thus cannot relate appropriately to entities that in significant ways are like us but very different otherwise. They would presumably be like us in terms of intelligence and ability to communicate. They would have different chemical compositions. But beyond that it is an open question now to what extent such machines would be like us. Crucially, we do not understand the nature of the mind and thus *what it takes to have one*. And so we do not know what to think of the prospect that machines might have minds. But what is nonetheless plausible is that some sophisticated machines must be accorded a moral status of sorts possibly to the point of human-rights protection.

This point gets much initial plausibility from noticing that, following James Moor, we can distinguish among kinds of moral status. To begin with, *ethical impact agents* are agents whose actions have ethical consequences whether intended or not. Any robot is such an agent if its actions can harm or benefit humans. Secondly, *implicit ethical agents* are agents whose design has ethical considerations built in, such as safety or security features. An example of a security feature is automatic teller machines checking availability or limiting the daily amount that can be withdrawn. *Explicit ethical agents* can secure and process

[29] On the challenges posted by a super-intelligence, see Bostrom, *Superintelligence*. See also Tegmark, *Life 3.0*.

[30] Corabi, "Superintelligence as Moral Philosopher."

ethical information about a variety of situations, make sensitive determinations about what to do, and even work out reasonable resolutions where considerations conflict. Finally, *full ethical agents* are explicit ethical agents who also have those central metaphysical features we usually attribute to agents *like us*, such as consciousness, intentionality or free will. [31]

In the space that opens between explicit and full ethical agents there could be a variety of different types of agency. Androids could display hallmarks of agency such as interactions with the environment combined with a level of independence and adjustability.[32] We may still not want to say explicit ethical agents falling short of full ethical agency deserve *all* the consideration full ethical agents deserve; but it would be equally implausible to say they deserve *none at all*. To get a better sense of what might be in store here let us explore how one might be able to argue in the first place that humans will remain the sole fully ethical agents. The crucial point would have to be that humans have a "mind" of sorts that machines cannot have, and that possession of such a mind merits a kind of protection (especially in terms of human rights) we would not have to grant to machines. Humans would be *conscious*, but machines would not be. But how would that be plausible?[33]

A traditional answer is that humans have *souls.* The general stance is metaphysical *substance dualism,* a set of views committed to the existence of non-physical mental

[31] Moor, "Four Kinds of Ethical Robots." For the field of machine ethics (which is concerned with giving machines ethical principles), see Wallach and Allen, *Moral Machines*. For an anthology on the subject, see Anderson and Anderson, *Machine Ethics*.

[32] Floridi and Sanders, "On the Morality of Artificial Agents." For an exploration of artificial morality and the agency of robots, see also Misselhorn, "Artificial Morality. Concepts, Issues and Challenges."

[33] (1) For the philosophy of mind behind what is to come, see Heil, *Philosophy of Mind*; Jaworski, *Philosophy of Mind*; Braddon-Mitchell and Jackson, *Philosophy of Mind and Cognition*; Carter, *Minds and Computers*. (2) Then there is the question of how much consciousness matters to begin with, see Levy, "The Value of Consciousness."

phenomena. This view is prominent in a range of religions, and philosophically is famously associated with Descartes. Not many contemporary philosophers defend such a view because of difficulties from accommodating mental substances within the worldview the natural sciences offer. Nonetheless, version of this view have defenders, and not merely among the religious. Some distinguished contemporary philosophers argue that *consciousness* is a primitive and basic component of nature. Thomas Nagel, for one, thinks "mind" cannot arise from physical substances and so must exist independently in nature in ways we do not yet understand. "In attempting to understand consciousness as a biological phenomenon," Nagel insists, "it is too easy to forget how radical is the difference between the subjective and the objective."[34]

But there is no reason to think that if indeed there are two types of substances in the world, machines would be categorically excluded from possessing both of them. How could we be certain that immensely sophisticated machines would not also host souls, if souls can be hosted at all? Or how could we be certain such machines could not host minds if consciousness exists independent in the world? It would be hard to fathom why the facts that we are made of carbon and reproduce sexually qualify us for possession of such mental substances in ways entities made from, say, silicon generated in non-sexual ways would not qualify. We have no conclusive reason to think either way at this stage.

[34] Nagel, *Mind & Cosmos*. The quote is from Nagel, 128. The emphasis on the differences between the subjective and the objective standpoint permeates Nagel's work, both his political philosophy and his philosophy of mind. In the philosophy of mind, he made the formulation of "what it's like to be something" central, see Nagel, "What Is It Like to Be a Bat?" (For the view that there could indeed be something it is like to be a robot and thus that robots can have a subjective point of view, see Kiverstein, "Could a Robot Have a Subjective Point of View?") For the political dimensions, see Nagel, *Equality and Partiality*. On the question whether machines can be conscious, see also Harnad, "Can a Machine Be Conscious? How?"

In addition to substance dualism there is *property dualism*, the view that the world is constituted of just one kind of substance - the physical kind – but there exist two distinct kinds of properties, physical and mental ones. Mental properties are neither identical with nor reducible to physical properties but may be instantiated by the same things that instantiate physical properties. Such views stir a middle course between substance dualism and physicalism. One version is *emergentism*, which holds that when matter is organized appropriately (the way living human bodies are organized), mental properties emerge in a way not accounted for by physical laws alone. In the contemporary debate, a form of this view has been espoused by David Chalmers.[35] Mental properties are basic constituents of reality on a par with fundamental physical properties such as electromagnetic charge. They may interact causally with other properties, but their existence does not depend upon any other properties. Here again there is no reason to think non-carbon material could not be organized in relevantly similar ways and thus give rise to the same properties.

So according to both versions of dualism it would simply be an open question whether eventually machines would have "minds" in whatever that would mean for humans. Another way of arguing that, plausibly, machines will differ from humans, however, comes from the physicalist side. Of one of the best-known contemporary philosophers who also operates as a public intellectual, Daniel Dennett is thoroughly physicalist in his outlook. Our understanding of ourselves includes not only the body and nervous system but, he argues, our consciousness with its elaborate sensory, emotional and cognitive features, as well as consciousness of other humans and nonhuman species. But Dennett thinks consciousness is

[35] Chalmers, *The Conscious Mind*.

a *user-illusion* indispensable for our dealings with each other and for managing ourselves. Our conception of conscious creatures with subjective inner lives allows us to predict how those creatures will behave. *Human* consciousness is to a large extent a product of cultural evolution involving memes, which generates minds rather different from those of other animals. Dennett coined the term "heterophenomenology" to describe the (scientifically false) attribution each of us makes to others of an inner representation of the world.[36]

But this also means that, once we see the full complexity of the brain, we understand it will be hard to create anything close to what it took evolution hundreds of millions of years to generate. Creating general AI, Dennett holds, is possible in principle, but "would cost too much and not give us anything we really needed."[37] Instead, we will tend to overestimate our abilities to construct such machines, prematurely ceding authority to them.[38] So on both Dennett's view and according to Heidegger as discussed earlier, it will be tremendously hard to imitate us. But on both views, the complexities involved in imitating us only support the point that something less sophisticated than us must already be accorded some kind of moral status. At least some explicit ethical agents will come up for serious moral consideration even if they fall short of being the same kind of full ethical agents that we are.

So far we have only discussed views in the philosophy of mind that seemingly make it harder to ascribe some kind of moral status to machines and found that none of them give us any reason to conclude that machines could never have any kind of moral status. There

[36] Dennett, *From Bacteria to Bach and Back*, chapter 14.

[37] Dennett, 400.

[38] Dennett, 402. See also Dennett, *Consciousness Explained*; Dennett, *Kinds Of Minds.* For a short introduction to Dennett's work and its importance to the philosophy of mind, see Heil, *Philosophy of Mind*, chapter 8. For more extensive discussion, see Brook and Ross, *Daniel Dennett*; Thompson, *Daniel Dennett*.

are other prominent views in the philosophy of mind that make it easier than those we just reported on to think machines can have some kind of moral status. Within the physicalist camp the (now largely defunct) mind/brain identity theory holds that states and processes of the mind are identical to states and processes of the brain. A successor to that view is *functionalism*, according to which what makes something a thought, desire, pain or any other type of mental state depends not on its internal constitution, but solely on what function or role it plays, in the cognitive system of which it is a part. Functionalism abstracts away from details of physical implementation by characterizing mental states in terms of non-mental functional properties. For example, a kidney is characterized scientifically by its functional role in maintaining chemical balances. It does not matter if kidneys are made up of organic tissue or silicon chips: it is the role they play and their relations to other organs that make them kidneys.

Functionalism's characterization of mental states in terms of their roles in the production of behavior grants them causal efficacy. Functionalism permits mental states to be *multiply realized*, and thereby offers an account of mental states compatible with physicalism without making brans and minds identical and thus without implausibly limiting the class of those with minds to creatures with brains like ours (as the original mind/brain identity theory had done). This is a theory much inspired by computer science: the mind relates to the brain like software relates to hardware. Software can run on different types of hardware, and similarly, very different types of physical entities can have minds.

Objections to functionalism include that a mind is too easy to come by in this way, and thus subjectivity is not suitably accommodated in this manner. In particular, there is John Searle's famous *Chinese Room argument*. Searle imagines himself alone in a room following

a computer program for responding to Chinese characters slipped under the door. He understands no Chinese, but by following the program for manipulating symbols just as a computer does, he produces appropriate strings of characters that fool those outside into thinking there is a Chinese speaker inside. The narrow lesson from this argument is supposed to be that programming a digital computer may make it appear to understand language but does not produce real understanding. The broader lesson is supposed to be a refutation of the theory that human minds are computer-like computational or information-processing systems.[39]

The ensuing rich debate in the philosophy of mind does not concern us. What matters is that for physicalists functionalism is the most prominent understanding of the mind, and in its terms the expectation that machines will have moral status is rather straightforward. That point is a straightforward consequence from the point that minds are multiply realizable.[40] So the upshot is that it is quite plausible that in the future machines will have

[39] For functionalism, see Heil, *Philosophy of Mind*, chapter 6. For an early formulation of functionalism, see Putnam, "Minds and Machines." The Turing test was also rather influential for this development, see Turing, "Computing Machinery and Intelligence." For influential critical discussion, see Block, "Troubles With Functionalism." For Searle's argument see Searle, "Minds, Brains and Programs." For an exchange among Seale, Chalmers and Dennett, see Searle, *The Mystery of Consciousness*. For functionalists, the following two principles proposed by Bostrom and Yudkowsky would be rather natural: (1) Principle of Substrate Non-Discrimination: If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status. (2) Principle of Ontogeny Non-Discrimination: If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status. See Bostrom and Yudkowsky, "The Ethics of Artificial Intelligence," 322–23.

[40] Relevant comparison to animal consciousness. Dawkins, Why Animals Matter, gives up on the question of consciousness and argues that taking the hard question of consciousness seriously brings us ever further away from coming to terms with animals. She thinks animals matter because animals need to be healthy humans to be healthy. Michael Tye's book is really good here – he does think we can conclude that animals are conscious in their own way and should then also be taken morally seriously. He also has an excellent chapter on machines. Consider here also PGS's book on octopus For a discussion of machine consciousness and ethical treatment of sophisticated machines in combination of a discussion of animals on the same issues, see Tye, *Tense Bees and Shell-Shocked Crabs*.

moral status of sorts, possibly at the level of human rights protection. Perhaps we should then pass a Universal Declaration of the Rights of Full Ethical Agents.

6. Conclusion

Scientific inquiry has moved us to the brink of revolutionizing technology to such an extent that AI is on our radar. But it is also precisely because we have made so much progress in understanding that we also increasingly grasp the limits of our understanding elsewhere. This is the broader understanding of the death-of-God-theme with which we started. Lack of agreement on moral foundations is a case in point. Another is our comprehension of consciousness. Challenge arise for the short, medium and long term, and it was the latter two that have concerned us here. The medium term is the period when technological innovation will have fully transformed society. The long term is when machines themselves come up for moral consideration. The distinction might be more analytical than chronological but is usefulness nonetheless.

Throughout we drew on insights from Heidegger's phenomenology. First of all, understanding technology it not merely as a set of tools but shapes the way we live and how we perceive fellow humans and other entities in the world. Secondly, there is much more than intelligence to a proper understanding of human life so that at least it will be much harder to substitute us than it would be if that were merely a matter of building intelligence. But this reassurance also entailed complexities, to wit, that "moral status" might be a much more complex matter in the future than it is now.

One major challenge about the medium term is how human rights will prevail in an era that quite possibly might be shaped by an enormous increase in economic inequality.

Human rights especially will be under siege. What is needed is greater focus on distributive justice, both domestically and globally, to make sure societies do not fall apart. And as far as this medium term is concerned, we must also heed the admonition we drew from Heidegger that technology must be used to enhance the distinctively human life rather than impoverishing it.

Even more than for the medium term we are woefully underprepared for the long-term challenges, in a way the death-of-God theme makes clear. We do not have settled views on when artificial minds would be sufficiently similar to human ones to be eligible for the same moral status. But we must be prepared to deal with more types of moral status than we currently do. And quite plausibly some machines will have some type of moral status, which may or may not fall short of the moral status of human beings. This is also a point emerging from the Heidegger discussion. Machines may have to be integrated into human social and political lives. We will be better able to do so if we integrate humans more properly in the medium run, through greater focus on distributive justice, both domestically and globally.

Bibliography

Anderson, Michael, and Susan Leigh Anderson, eds. *Machine Ethics*. 1 edition. New York: Cambridge University Press, 2011.

Bittner, Rüdiger. *Bürger sein: Eine Prüfung politischer Begriffe*. 1 edition. Berlin Boston: De Gruyter, 2017.

Blackmore, Susan. *Conversations on Consciousness: What the Best Minds Think about the Brain, Free Will, and What It Means to Be Human*. 1 edition. Oxford; New York: Oxford University Press, 2007.

Blackmore, Susan, and Emily T. Troscianko. *Consciousness: An Introduction*. 3rd New edition edition. Abingdon, Oxon ; New York, NY: Routledge, 2018.

Block, Ned. "Troubles With Functionalism." In *Readings in the Philosophy of Psychology, Volumes 1 and 2,* 268–305. Cambridge, Mass: Harvard University Press, 1980.

Borgmann, Albert. *Technology and the Character of Contemporary Life: A Philosophical Inquiry*. Chicago: The University of Chicago Press, 1987.
Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford, United Kingdom ; New York, NY: Oxford University Press, 2016.
Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 316–34. Cambridge, UK: Cambridge University Press, 2014.
Braddon-Mitchell, David, and Frank Jackson. *Philosophy of Mind and Cognition: An Introduction*. 2 edition. Malden, MA: Wiley-Blackwell, 2006.
Brook, Andrew, and Don Ross, eds. *Daniel Dennett*. 1st edition. Cambridge ; New York: Cambridge University Press, 2002.
Carman, Taylor. *Merleau-Ponty*. 1 edition. London ; New York: Routledge, 2008.
Carter, Matt. *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence*. 1 edition. Edinburgh: Edinburgh University Press, 2007.
Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. 1st edition. New York: Oxford University Press, 1996.
———. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, no. 9–10 (2010): 7–65.
Corabi, Joseph. "Superintelligence as Moral Philosopher." *Journal of Consciousness Studies* 24, no. 5 (2017): 128–49.
Dennett, Daniel C. *Consciousness Explained*. 1 edition. Boston: Back Bay Books, 1992.
———. *From Bacteria to Bach and Back: The Evolution of Minds*. 1 edition. New York: W. W. Norton & Company, 2018.
———. *Kinds Of Minds: Toward An Understanding Of Consciousness*. 4th printing edition. New York: Basic Books, 1997.
Dietsch, Peter. *Catching Capital*. New York, NY: Oxford Univ Pr, 2015.
Dreyfus, Hubert. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York City: Harper & Row, 1972.
Dreyfus, Hubert L. *On the Internet*. 2 edition. Milton Park, Abingdon, Oxon ; New York, NY: Routledge, 2008.
———. *What Computers Still Can't Do: A Critique of Artificial Reason*. Revised ed. edition. Cambridge, Mass: The MIT Press, 1992.
Espinet, David, Günter Figal, Tobias Keiling, and Nikola Mirkovic. *Heideggers "Schwarze Hefte" im Kontext: Geschichte, Politik, Ideologie*. 1st ed. Tübingen: Mohr Siebeck, 2018.
Floridi, Luciano, and J. W. Sanders. "On the Morality of Artificial Agents." *Mind and Machine* 14 (2004): 349–79.
Harnad, Steve. "Can a Machine Be Conscious? How?" *Journal of Consciousness Studies* 10, no. 4–5 (2003): 67–75.
Haugeland, John. *Artificial Intelligence: The Very Idea*. First edition. Cambridge, Mass: MIT Press, 1985.
Heidegger, Martin. *Being and Time*. Reprint edition. New York: Harper Perennial Modern Classics, 2008.
———. *The Question Concerning Technology, and Other Essays*. Reissue edition. New York; London Toronto: Harper Perennial Modern Classics, 2013.

Heil, John. *Philosophy of Mind: A Contemporary Introduction*. 3 edition. New York, NY: Routledge, 2012.
Jaworski, William. *Philosophy of Mind: A Comprehensive Introduction*. 1 edition. Chichester, West Sussex ; Malden, MA: Wiley-Blackwell, 2011.
Kiverstein, Julian. "Could a Robot Have a Subjective Point of View?" *Journal of Consciousness Studies* 14, no. 7 (2007): 127–39.
Kiverstein, Julian, and Michael Wheeler, eds. *Heidegger and Cognitive Science*. 2012 edition. New York: Palgrave Macmillan, 2012.
Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books, 2006.
Levy, Neil. "The Value of Consciousness." *Journal of Consciousness Studies* 21, no. 1–2 (2014): 127–38.
McLellan, David. *Karl Marx: Selected Writings*. Oxford: Oxford University Press, 1977.
Misselhorn, Catrin. "Artificial Morality. Concepts, Issues and Challenges." *Society* 55 (2018): 161–69.
Mitchell, Andrew J., and Peter Trawny, eds. *Heidegger's Black Notebooks: Responses to Anti-Semitism*. New York: Columbia University Press, 2017.
Moor, James H. "Four Kinds of Ethical Robots." *Philosophy Now* 72 (2009): 12–14.
Moyn, Samuel. *Not Enough: Human Rights in an Unequal World*. Cambridge, Massachusetts: Belknap Press: An Imprint of Harvard University Press, 2018.
Nagel, Thomas. *Equality and Partiality*. 1 edition. New York: Oxford University Press, 1991.
———. *Mind & Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. 1st edition. New York: Oxford University Press, 2012.
———. "What Is It Like to Be a Bat?" *Philosophical Review* 83, no. 4 (1974): 435–50.
Nietzsche, Friedrich. *The Gay Science: With a Prelude in Rhymes and an Appendix of Songs*. Translated by Walter Kaufmann. 1 edition. New York: Vintage, 1974.
Noë, Alva. *Out of Our Heads: Why You Are Not Your Brain, and Other Lessons from the Biology of Consciousness*. 1 edition. New York: Hill and Wang, 2010.
Olafson, Frederick A. *Heidegger and the Philosophy of Mind*. 1st Edition edition. New Haven: Yale University Press, 1987.
Piketty, Thomas. *Capital in the Twenty-First Century*. Cambridge: Belknap, 2014.
Putnam, Hilary. "Minds and Machines." In *Mind, Language, and Reality*, 362–385. Cambridge: Cambridge University Press, 1975.
Raso, Filippo, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. "Artificial Intelligence & Human Rights: Opportunities & Risks." Berkman Klein Center for Internet & Society at Harvard University, 2018.
Rawls, John. *A Theory of Justice*. Cambridge: Harvard University Press, 1971.
Richardson, John. *Heidegger*. 1 edition. New York: Routledge, 2012.
Risse, Mathias. *On Global Justice*. Princeton: Princeton University Press, 2012.
Risse, Mathias, and Gabriel Wollner. *On Trade Justice: A Philosophical Plea for a New Global Deal*. Oxford: Oxford University Press, under contract.
Scheidel, Walter. *Great Leveler: Violence and the History of Inequality from the Stone Age to the Twenty-First Century*. Princeton, NJ: Princeton Univers. Press, 2017.
Searle, John. "Minds, Brains and Programs." *Behavioral and Brain Sciences* 3 (1980): 417–57.

Searle, John R. *The Mystery of Consciousness*. 1st edition. New York: The New York Review of Books, 1997.
Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.
Thompson, David L. *Daniel Dennett*. 1 edition. London ; New York: Continuum, 2009.
Turing, Alan. "Computing Machinery and Intelligence." *Mind* 59, no. 236 (1950): 433–60.
Tye, Michael. *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* 1 edition. New York, NY: Oxford University Press, 2016.
Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. 1 edition. Oxford: Oxford University Press, 2010.
Wolin, Richard. *The Politics of Being: The Political Thought of Martin Heidegger*. With a new preface edition. New York City, NY: Columbia University Press, 2016.
Wrathall, Mark, and Simon Critchley. *How to Read Heidegger*. 1 edition. New York: W. W. Norton & Company, 2006.
Young, Julian. *Heidegger's Later Philosophy*. Cambridge: Cambridge University Press, 2001.

Carr Center for Human Rights
Policy Harvard Kennedy School
79 John F. Kennedy Street
Cambridge, MA 02138

www.carrcenter.hks.harvard.edu