The Future Impact of Artificial Intelligence on Humans and Human Rights

Steven Livingston and Mathias Risse

In recent years there has been an enormous increase in reflection on the impact of artificial intelligence (AI) on society and politics. It is now often computer scientists who admonish the rest of us that impact assessments are far ahead of what current engineering can deliver. But what engineering can deliver now may offer little indication of how long it will take for technological innovation to revolutionize this or that sector of society, and thus ultimately society as a whole. The overall pace of innovation has picked up over the past few decades, and today innovations spread with breathtaking speed. Consequently, there are many questions that should be on our radar now even though they do not currently arise with any acuteness—and it will be too late to start thinking about them only once they do. The human rights domain is no exception here.

This essay begins with a brief introduction to AI. We then offer examples of the way AI has already had an effect on human rights monitoring practices. Next we turn to an exploration of artificial general intelligence (AGI), also called superintelligence. With that groundwork laid, we discuss what humans might owe superintelligent machines. Indeed, machines may come up for moral status even before there is an actual superintelligence. We then address the reverse question of what such machines would owe humans and how we would manage the potential collapse of the binary distinction between machine and human. Finally, we explain why we cannot offer real conclusions.[1]

Ethics & International Affairs, 33, no. 2 (2019), pp. 141–158.
© 2019 Carnegie Council for Ethics in International Affairs
doi:10.1017/S089267941900011X

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

The Nature of Artificial Intelligence

"Artificial Intelligence" is often used interchangeably with "machine learning" and "deep learning." With machine learning, algorithms are trained to identify relationships, develop predictive models, and make decisions. The better the training data, the better the algorithm. One example is image classification. Here, machine learning involves *manually selecting* the relevant features of an image, such as the edges and corners of objects, and using this information to create a model to identify those objects. In some cases, users of websites, social media, and other platforms train algorithms by interacting online, buying various items, viewing content, and posting and liking content.[2]

Deep learning also uses training data to find patterns used to make predictions about new data, though in this case features are not extracted by humans. Rather, data are fed directly into the deep learning algorithm, which then predicts the occurrence of objects. Deep learning algorithms do this via multiple layers of artificial neural networks that mimic biological brains. Output data of one layer is input data for the next layer, and so forth. Deep learning architecture is used in speech recognition, natural language processing, and computer vision.[3]

Finally, "reinforcement learning" focuses on improving the performance of an algorithm over multiple engagements with a problem, adjusting actions based on continuous feedback from past actions. While deep learning and reinforcement learning are not mutually exclusive, what is most significant about reinforcement learning is the possible elimination of training data. It unshackles AI from human limitations. We will describe this in greater detail later in this essay when we review an algorithm called AlphaZero.

The terms "general AI," "AGI," and "superintelligence" are often used interchangeably. AGI refers to an algorithm or set of algorithms that are, minimally, as capable as a typical human across multiple problem domains. Importantly, *as of this writing AGI does not exist.* Yet several organizations are making rapid advances in expanding the adaptability of algorithms to multiple problem domains.[4] Google's DeepMind and Google Brain, along with initiatives at other major tech companies, are pushing AI toward superintelligence. Greg Brockman, co-founder and chief technology officer at the company OpenAI, recently noted that after fifty or more years of research, AGI now seems likely in the near term.[5]

The central point of our essay is to help us to think about the implications of AGI for humans. After a brief review of some of the effects that machine

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

and deep learning have had on current human rights practices, we return to a discussion of AGI.

AI's Current Impact on Human Rights

So far, standard machine learning has offered an improved ability to monitor and document war crimes and human rights abuses. Digital technology in the twenty-first century has ushered in what some have called the "golden age of surveillance"—not only by states and corporations but also by nonstate actors.[6] Human rights groups, news organizations, and open-source investigators such as Bellingcat and the Syrian Archive access massive amounts of open-source data generated by billions of sensor platforms in the hands and pockets of people around the globe.

With the proliferation of multipurpose mobile phones and other imaging platforms, including hundreds of high-resolution imaging satellites, something approaching ubiquitous surveillance has emerged. If something happens, almost anywhere in the world, there is a good chance it will be recorded by a camera either on the ground or in orbit. In 2017, for instance, the International Criminal Court issued an indictment for the arrest of a Libyan warlord based on satellite imagery and videos taken of the executions he ordered (or conducted himself) and that were posted to social media by his followers.[7] Geographical features seen in the videos—buildings, roads, trees, hills—were located via time-stamped high-resolution satellite images.[8] In this way, video, photos, satellite images, and other data are triangulated to verify events in a specific time and place. In the case of the Libyan warlord, much of the analysis was done by humans engaged in many hours of painstaking review of satellite imagery.

Machine learning is automating this sort of work. Planet, one of many satellite data analytics firms, runs Planet Analytics, a set of machine learning algorithms that help customers detect and classify geographical features and monitor subtle changes over time.[9] Similarly, a Google application called PlaNet uses deep learning to geolocate landscape photographs captured anywhere on Earth.[10] After subdividing the surface of the Earth into millions of multiscale geographical cells, Google trained a neural network to locate the place on the Earth's surface where an image is taken.[11] Similarly, in a 2016 effort to map the global population who do not have Internet connectivity, Facebook analyzed 14.6 billion geospatial images capturing some 21.6 million square kilometers of the Earth's surface using the same AI pattern recognition tool it uses for facial recognition.[12]

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

AI has also transformed another field important to human rights investigations. Forensic anthropology has played a significant role in human rights abuse documentation since the 1980s, involving the examination of bones and other physical evidence in an effort to reconstruct the circumstances of death. In recent years, DNA sequencing has introduced a much greater degree of scientific accuracy and efficiency in forensic investigations. Even scattered bone fragments offer genetic evidence that can establish the identity of victims. What is known as "massively parallel DNA sequencing" makes use of AI to sequence strands of DNA more accurately, more efficiently, and at a lower cost.[13] In general, machine learning and deep learning allow human rights groups to discover evidence that would otherwise be unavailable. If this trend were to continue, we would expect a much higher degree of awareness of war crimes and abuse around the world.

Yet technologies rarely follow tightly prescribed paths to specific effects. Human ingenuity and cultural practices bend them toward unexpected outcomes. This idea is captured by the notion of *affordances*—the range of possible outcomes enabled by a technology's design features. A chair, for example, enables sitting, though it can also be used as a coat rack, a stepladder, or a door jam.[14] Some software platforms allow for only reading (a website, for example) while others invite reading *and* content creation (sites such as Twitter, Facebook, and Instagram). Politically, one affordance invites hierarchical propagation of information while the other invites participatory expressive acts.[15] Human intentions mediate particular design outcomes.[16] The AI capacities used to track warlords or identify the missing are also used by authoritarian states to monitor citizens.[17] In China, the social credit score system uses ubiquitous observational surveillance, facial recognition software, and social media surveillance to track citizen behavior and assign scores according to how well people comply with rules large and small.[18]

Artificial neural networks can uncover patterns in massive amounts of data that lead to the discovery of evidence of a war crime, or they can be configured to create "deepfakes"—an AI application that seamlessly alters visual content. They can be used for processes like digitally inserting the face and voice of one person into a video of another person, creating highly realistic but entirely synthetic video content. When seeing is no longer believing, the epistemological foundation of journalism, human rights investigations, judicial proceedings, and other fact-based processes is jeopardized. Rather than serving as a source of clarification, AI is used here to delude, obfuscate, and humiliate targeted individuals and

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

organizations.[19] Specific outcomes of current AI affordance are difficult to predict. What might the future hold in store?[20]

Artificial General Intelligence

Let us consider a twenty-five-year-old prediction made about the effects of AI on society. At the 1993 VISION-21 Symposium sponsored by the NASA Lewis Research Center and the Ohio Aerospace Institute, computer scientist Vernor Vinge predicted that "within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will end." In making this claim, Vinge was not suggesting that there would be some cyborg apocalypse. Instead, humans would remain, but in a radically transformed condition. Not only might something like superintelligence be realized but "computer/human interfaces may become so intimate that users may reasonably be considered superhumanly intelligent."[21] This would, he said, signal the end of the human era, as the new superintelligence would advance at an incomprehensible rate. Such an AI knowledge cascade is sometimes called the "singularity." According to this hypothesis, a self-improving, self-learning intelligent agent would enter a cycle of improvement based on a "runaway reaction." At such a moment, the AI would quickly surpass all human intelligence.[22] Vinge offers relatively clear markers for assessing the current state of AI. But what evidence is there that Vinge's prediction will be realized?

As of this writing, the algorithms of DeepMind Technologies present what is probably the best case for near-term realization of AGI.[23] Public awareness of DeepMind's computer program AlphaGo grew enormously in 2016 after it defeated eighteen-time Go world champion Lee Sedol four games to one. An ancient board game immensely popular in Korea, Japan, and China, Go is played on a square 19 x 19 grid board and has 10 to the 171st power possible legal positions. It is regarded as the most complex board game in existence. Using labelled training data consisting of thirty million moves taken from 160,000 games by top-ranked human players, AlphaGo was trained to mimic the best human players in the world.[24] Subsequent iterations of the Alpha algorithm rely on reinforcement learning done by the program, rather than training data. Reinforcement learning takes action to maximize reward according to specified objectives. Whereas supervised learning—as used in training AlphaGo—relies on labelled training data (the thirty million expert moves), reinforcement learning involves only the algorithm

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

playing against previous versions of itself thousands of times, learning and improving with each cycle of play.[25] The first subsequent iteration was known as AlphaGo Zero. Newer iterations have followed, all with different names. As computer scientist David Silver and his colleagues describe it, "In each iteration, the performance of the system improves by a small amount, and the quality of the self-play games increases, leading to more and more accurate neural networks and ever stronger versions of AlphaGo Zero."[26]

In three hours of reinforcement learning, AlphaGo Zero had achieved the competency of a human beginner; in nineteen hours, it had reached the level of an advanced player; in three days, it achieved world-class capabilities and defeated AlphaGo—the same version that had just months before defeated the best human player in the world—one hundred games to none.[27] And it accomplished this by eliminating the weakest element in the algorithm's training: *humans*.[28]

In 2018, DeepMind applied its AI capacities to biological research in the form of the AlphaFold algorithm.[29] In a protein structure prediction contest among teams of scientists, AlphaFold successfully predicted twenty-five of forty-three protein structures. The second-place human team predicted only three.[30] But is this superintelligence or AGI? Dr. Oren Etzioni, chief executive officer of the Allen Institute for Artificial Intelligence, observed,

> I think it would be a mistake to believe that we've learned something general [from DeepMind] about thinking and about learning for general intelligence. *This approach won't work in more ill-structured problems like natural-language understanding or robotics, where the state space [the set of all possible configurations of a system] is more complex and there isn't a clear objective function.*[31]

A "state space" is the set of all possible configurations of a system.

DeepMind cofounder Demis Hassabis does not claim that AlphaGo experiences consciousness, though he has said it is creative and intuitive. A research scientist at DeepMind, Razvan Pascanu, and his colleagues note that "imagining the consequences of your actions before you take them," as the AlphaGo algorithms do, "is a powerful tool of human cognition. If our algorithms are to develop equally sophisticated behaviors, they too must have the capability to 'imagine' and reason about the future."[32] In January 2019, DeepMind unveiled AlphaStar to compete in a series of competitive StarCraft computer games.[33] StarCraft leans further in the direction of the complex state space Etzioni pointed to in his earlier observations.

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

AlphaStar defeated professional StarCraft players in ten of eleven games.[34] DeepMind seems to be making progress in the direction of Vinge's vision of 2023.

Recall that Vinge said that by 2023 "computer/human interface interfaces may become so intimate that users may reasonably be considered superhumanly intelligent."[35] What is the current state of technology in terms of computer-human interfaces?

Putting aside the argument that the obsessive use of smartphones and other devices might in itself meet Vinge's threshold for an intimate computer-human interface, we can see the near-term development of other technologies that come even closer. Neuralink, an American neurotechnology company founded in 2016 by Elon Musk, is developing implantable brain-computer interfaces intended to link cognitive capacity to the Internet. The company's initial goal is to create a "'digital layer' located above the cortex."[36] Already, neural implants are used by people with Parkinson's disease to steady tremors. Epilepsy patients also have implanted electronic monitors to detect signs of impending seizures and emit electric pulses to prevent them. The long-term goal for Neuralink, though, is a "symbiosis with artificial intelligence."[37] And Musk is not alone in his quest to develop a neural link, as Vinge envisioned in 1993. Arnav Kapur, a graduate student and research assistant at MIT's Media Lab, developed AlterEgo, a device that is worn around one's ear to monitor internal thoughts. Electrical signals that the brain normally sends to the vocal cords are intercepted and redirected to a computer and the Internet. The computer response is then communicated to the user's inner ear via vibrations.[38] This entire cycle is measured in a fraction of a second.

Vinge imagined humans would, in a sense, disappear into their devices. For Vinge (and Musk), the connection would be seamless and wired directly to cognitive function without the encumbrances of hands, fingers, and thumbs to slow things down. As DeepMind and OpenAI at least approach AGI, and as Musk and others look to alter the link between humans and computers, we are nearing a moment not all that unlike what Vinge imagined. What are the ethical implications of these developments? What do we owe intelligent machines and what do intelligent machines owe us? In a blended world of machine and human, does drawing a distinction even make sense?[39]

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

AI and Evolving Notions of Moral Status

It is plausible that at some point in the not too distant future intelligent machines will be accorded a moral status, even approximating that of a human rights protection. This point gets much initial plausibility from noticing that, following James Moor, we can distinguish among kinds of moral status.[40] 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. Second, "implicit ethical agents" are those whose designs have ethical considerations built in, such as safety or security features. An example of a security feature is an ATM checking the availability of funds and/or limiting the daily amount that an account holder can withdraw. "Explicit ethical agents" can secure and process 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 only to human agents, such as consciousness, intentionality, and free will.

In the space between explicit and full ethical agents, there could be a variety of different types of agency. AGI agents could display hallmarks of ethical agency, such as interactions with the environment combined with a level of independence and adjustability.[41] We may still not want to say that explicit ethical agents that fall short of full ethical agency deserve all the consideration that 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 entities that are full ethical agents. The crucial point would have to be that humans have a "mind" of the sort 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 grant machines. This is a claim that exceeds anything like DeepMind's claims to AGI. Humans would be conscious but machines would not be, even if they show imagination. But how would that be plausible?[42]

A traditional answer is that humans have souls. The general stance is substance dualism, a set of views in metaphysics committed to the existence of nonphysical mental phenomena. Not many contemporary philosophers defend such a view because of difficulties that arise from trying to accommodate mental substances within the worldview the natural sciences offer. Nonetheless, versions of this

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

view have defenders, and not merely among the religious. Some distinguished philosophers argue that consciousness is a primitive and basic component of nature. Thomas Nagel, for one, thinks the 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."[43]

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 like DeepMind would not at some point 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 independently in the world? It is hard to fathom why the fact that we are made of carbon and reproduce sexually qualifies us for possession of such mental substances in ways entities made from, say, silicon and generated in nonsexual ways would not qualify. We have no conclusive reason to think either way at this stage.

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 that there exist two distinct kinds of properties: physical and mental. 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 steer 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.[44] 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 that noncarbon material and, plausibly, also processes such as those involved with the DeepMind algorithm 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 will have minds. Another way of plausibly arguing that machines will differ from humans, however, comes from the physicalist

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

side. 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 also, he argues, our consciousness with its elaborate sensory, emotional, and cognitive features, as well as awareness of other humans and nonhuman species. Dennett thinks consciousness is a user illusion indispensable for our dealings with one another and for managing ourselves. Our conception of conscious creatures with subjective inner lives allows us to predict how we as creatures will behave. Human consciousness is, on this understanding, to a large extent a product of cultural evolution involving memes, a process that generates minds different from those of other animals.[45]

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 AGI, Dennett holds, is possible in principle, but "would cost too much and not give us anything we really needed."[46] Instead, we will tend to overestimate our abilities to construct such machines, prematurely ceding authority to them. It will be tremendously hard to imitate human life. Of course, DeepMind's AlphaFold and AlphaStar raise questions about the long-term viability of this position. But that there are so many complexities involved in imitating humans only serves to 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 humans are.

Thus far, we have only discussed views in the philosophy of mind that, on the face of it, would make it hard to ascribe moral status to machines, none of which gives us any reason to conclude machines could never be granted moral status. There are other prominent views that make it easier to think machines can have moral status related, if not identical, to that of humans. Among philosophers who believe physics fully describes the universe, the most prominent understanding of the mind is "functionalism." This is a theory much inspired by computer science: Roughly, the mind relates to the brain as software relates to hardware. Software can run on different types of hardware and, similarly, different types of physical entities can have minds. This is the philosophical home most welcoming to claims of AGI and superintelligence as a kind of conscious mind. What matters most for us is that functionalism permits mental states to be multiply realized—in entities composed of carbon as well as silicon. For physicalists, again,

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

functionalism is the most prominent understanding of the mind. On its terms, the expectation that machines can have moral status is straightforward.[47]

The upshot is that it is plausible that in the future machines will have moral status of sorts, possibly at the level of human rights protection. Perhaps we would then have to replace the Universal Declaration of Human Rights with the "Universal Declaration of the Rights of Full Ethical Agents." But let us now turn the question around and ask what rights humans might be able to claim in a world of AGI—and possibly superintelligence. Before doing so, we will examine the collapse of the binary distinction between machine and human.

AI and Evolving Notions of the Human

The question of the moral status of machines is made more complex when we abandon the dualism between humans and machines. Liberalism and human rights are predicated on the existence of individual human beings endowed with moral claims. If, however, Musk's vision of a neural link becomes a reality, as it seems on course to do, where does the individual begin and the machine end? As Yuval Noah Harari has noted, "For liberalism to make sense, I must have one—and only one—true self, for if I had more than one authentic voice, how would I know which voice to heed in the polling station, in the supermarket, and in the marriage market?"[48] The emergence of a human-machine mélange undercuts liberal assumptions about individualism. Even today, without the benefit of a direct neural link between the Internet and human cognition, scholars see the erosion of individual free will in market decisions made in the face of highly manipulative, individually tailored appeals based on thousands of data points—gathered from social media posts, location data, demographic data, and dozens of other indicators—and analyzed by deep-learning neural networks. Shoshana Zuboff calls this "surveillance capitalism."[49] The point is that even without neural links, the Internet and deep learning are shaping consciousness.

But who needs consciousness? Skeptics say that even the most advanced AGI will not achieve consciousness, which it would need to do in order to reach and exceed human intelligence. Harari responds by noting,

> There might be several alternative ways leading to super-intelligence, only some of which pass through the straits of consciousness. For millions of years organic evolution has been slowly sailing along the conscious route. The evolution of inorganic computers

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

> may completely bypass these narrow straits, charting a different and much quicker course to superintelligence. This raises a novel question: which of the two is really important, intelligence or consciousness? As long as they went hand in hand, debating their relative value was just an amusing pastime for philosophers. But in the twenty-first century this is becoming an urgent political and economic issue. And it is sobering to realise that, at least for armies and corporations, the answer is straightforward: intelligence is mandatory but consciousness is optional.[50]

Indeed, from some perspectives, such as military planning, consciousness might even be suboptimal.

AI might very well enhance humans in even more fundamental ways. In addition to humans tapping into AI, AI might redesign humans. By altering DNA sequences and modifying gene function, an AI-enabled gene editing tool called CRISPR (clustered regularly interspaced short palindromic repeats) allows researchers to treat certain chronic diseases, which is of course wonderful news to those who suffer from them.[51] As salubrious as the emerging technology can be, gene editing also offers the unsettling possibility of creating superhumans with heightened resistance to disease, greater intelligence, and increased physical endurance.[52] Again, as Harari notes, "Homo sapiens is not going to be exterminated by a robot revolt. Rather, Homo sapiens is likely to upgrade itself step by step, merging with robots and computers in the process . . . . In pursuit of health, happiness and power, humans will gradually change first one of their features and then another, and another, until *they will no longer be human*."[53] The process will most likely unfold gradually, with the most well-resourced and politically connected sectors of society availing themselves and their offspring of the advantages of bioengineering in much the same way they currently avail themselves and their offspring of the advantages of better healthcare, schools, housing, and leisure time. The trends in wealth inequality identified by Thomas Piketty will produce even starker distinctions between the haves and the have-nots.[54]

AI pessimists like Elon Musk claim the only hope for human survival in the face of inevitable AI superintelligence is to be found in inferior humans merging with advanced machines. Musk's solution, however, is fraught with ethical problems. Assuming that not all 9.7 billion persons expected to inhabit the planet in 2050 will have access to superintelligence through a neural link, or access to gene editing, what rights will be afforded to the billions of unenhanced humans left behind?[55] What ethical obligations will enhanced humans have to the unenhanced? What obligations will an AI superintelligent agent have to *any* human,

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

enhanced or otherwise? Max Tegmark remarks, "If we create AI that is smarter than us we have to be open to the possibility that we might actually lose control to them."[56] If AI superintelligence emerges, it is not readily obvious why it would tolerate humans, much less human rights.[57]

Superintelligent AI might also lead to a shift in political authority.[58] In the pre-Enlightenment era political authority was ascribed to god(s), often thought to dwell in the clouds far above the mortal realm. With the emergence of humanism and the Enlightenment, authority claims shifted to human reason, evidence, and institutions.[59] "We the People" became the source of political authority and knowledge claims. An emerging third era is returning authority to the clouds, though this time it is cloud computing. With superintelligent AI, human claims to authority in decision-making are weakening in problem domains as diverse as travel directions and criminal sentencing.[60] Further, it is unclear what claims to authority human institutions will have over enhanced humans defined by their direct cognitive access to superintelligent AI. One suspects that, like the global elite today, in most cases enhanced humans will live beyond the reach of conventional political authority.[61]

The Morality of Pure Intelligence

Philosophers have, of course, long debated the relationship between morality and rationality. If indeed rationality alone generates certain moral obligations, then presumably we could rest assured that superintelligent (and hence presumably rational) machines would live up to these obligations in such ways as highly imperfect humans never could.[62] Most famously perhaps is the dispute between David Hume and Immanuel Kant about whether rationality fixes our values. Hume thought reason does nothing to fix values: any being (including, for our purposes, AGI) endowed with reason, rationality, or intelligence (let us indeed assume these are all relevantly similar) might have any goals, as well as any range of attitudes, especially toward humans.[63] If so, an AGI or superintelligence—or any AI for that matter—could have just about any type of value commitment, including ones that might be extremely detrimental to human rights. The Kantian view, however, derives morality from rationality.[64] Kant's categorical imperative asks of all rational beings that they not ever use their own rational capacities nor those of any other rational being in a purely instrumental way. Excluded, in particular, are gratuitous violence against and the deception of

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

other rational beings.[65] The point of Kant's derivation is that any intelligent being would fall into a contradiction with itself by violating other rational beings. Roughly speaking, that is because it is only our rational choosing that gives any value to anything in the first place, which also means by valuing anything at all we are committed to valuing our capacity to value. If Kant is right, a superintelligence might be a true role model for ethical behavior. In this case, AI might close the gap that opens when humans, with their stone-age, small-group-oriented DNA, operate in a global context, helping to relieve us of our parochial judgments and failings.[66]

The Kantian model assumes that we would be rational *enough* for this kind of argument to generate protection for humble humans in an era of much smarter machines. It is similar in some important respects to T. M. Scanlon's ideas about appropriate responses to values.[67] The superintelligence might be "moral" in the sense of reacting in appropriate ways toward what it observes all around. On the one hand, perhaps under such conditions we would have some chance at earning the respect of superintelligent machines, given that the abilities of the human brain are truly astounding.[68] On the other, so are the abilities of animals, but that has not normally led humans to react toward them (or the environment) in an appropriately respectful way. Instead of displaying something like an enlightened anthropocentrism, we have too often instrumentalized nature. Hopefully, a superintelligence would simply outperform us in such matters, such that distinctively human life will receive some protection because we are worthy of respect. We cannot know for sure, but we also need not be wholly pessimistic.

Another interesting possibility is that in a world with multiple or even many AGI agents, a Hobbesian state of nature would apply to the original status of superintelligences vis-à-vis each other, such that they would eventually subject themselves to some kind of shared higher authority—an AI leviathan of sorts. Whether such a shared authority would also create benefits for humans is unclear.[69]

Why There is No Real Conclusion

We started by looking at some of the many ways in which AI has already had an impact on human rights. We then turned to an exploration of AGI with a particular interest in how close we are to its development. With that exploration in

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

place, we pursued two major questions. First, what do humans owe superintelligent machines? We argue that machines too may have moral status, even before there is an actual superintelligence. And second, what would such machines owe humans and how would we manage the potential collapse of the binary distinction between human and machine? We have little to list here by way of actual conclusions because these questions are still so new. AI's relevance to human rights is just beginning to dawn on us. What is enormously likely, however, is that these questions will take on ever greater relevance over the next few decades.

NOTES

[1] For the nexus between human rights and AI, see Mathias Risse, "Human Rights and Artificial Intelligence: An Urgently Needed Agenda," *Human Rights Quarterly* 41, no. 1 (February 2019), pp. 1–16.
[2] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Human Power* (New York: PublicAffairs, 2019).
[3] For a more complete description of deep learning, see Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge, Mass.: MIT Press, 2017), www.deeplearningbook.org/.
[4] Graham Webster, Rogier Creemers, Paul Triolo, and Elsa Kania, "China's 'New Generation Artificial Intelligence Development Plan' (2017)," *New America* blog, July 20, 2017, www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.
[5] "Can We Rule Out Near-Term AGI?," YouTube video, 17:19, posted by Web Summit, November 7, 2018, www.youtube.com/watch?time_continue=1&v=YHCSNsLKHfM. For a contrary view, see Trent Eady, "Does Recent Progress with Neural Networks Foretell Artificial General Intelligence?," Medium, December 4, 2018, medium.com/protopiablog/does-recent-progress-with-neural-networks-foretell-artificial-general-intelligence-9545c17a5d8b.
[6] Regarding state surveillance, see Ronald J. Deibert, "The Road to Digital Unfreedom: Three Painful Truths about Social Media," *Journal of Democracy* 30, no. 1 (January 2019), pp. 25–39; regarding corporate surveillance, see Zuboff, *Age of Surveillance Capitalism*; regarding open-source investigations by nonstate actors, see Steven Livingston and Sushma Raman, "Human Rights Documentation in Limited Access Areas: The Use of Technology in War Crimes and Human Rights Abuse Investigations" (Cambridge, Mass.: Carr Center for Human Rights Policy, May 2018), carrcenter.hks.harvard.edu/files/cchr/files/documentationandtech_designed_may_8_2018.pdf.
[7] Peter Cluskey, "Social Media Evidence a Game-Changer in War Crimes Trial," *Irish Times*, October 3, 2017, www.irishtimes.com/news/world/europe/social-media-evidence-a-game-changer-in-war-crimes-trial-1.3243098.
[8] Bellingcat Investigation Team, "How a Werfalli Execution Site Was Geolocated," Bellingcat, October 3, 2017, www.bellingcat.com/news/mena/2017/10/03/how-an-execution-site-was-geolocated/.
[9] Shawna Wolverton, "Making the Move from Imagery to Insights with Planet Analytics," Planet, July 18, 2018, www.planet.com/pulse/planet-analytics-launch/.
[10] A. J. Rohn, "Google's PlaNet: Geolocating Photos Using Artificial Intelligence," GIS Lounge, March 11, 2016, www.gislounge.com/google-planet-geolocating/.
[11] Tobias Weyand, Ilya Kostrikov, and James Philbin, "PlaNet — Photo Geolocation with Convolutional Neural Networks," arXiv.org, submitted February 17, 2016, arxiv.org/pdf/1602.05314.pdf.
[12] Clay Dillow, "What Happens When You Combine Artificial Intelligence and Satellite Imagery," *Fortune*, March 30, 2016, fortune.com/2016/03/30/facebook-ai-satellite-imagery/.
[13] Edo D'Agaro, "Artificial Intelligence Used in Genome Analysis Studies," *EuroBiotech Journal* 2, no. 2 (April 2018), www.degruyter.com/downloadpdf/j/ebtj.2018.2.issue-2/ebtj-2018-0012/ebtj-2018-0012.pdf.
[14] Donald Norman, *The Design of Everyday Things* (New York: Basic Books, 1988).
[15] W. Lance Bennett and Alexandra Segerberg, *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics* (New York: Cambridge University Press, 2013).
[16] James J. Gibson, *The Ecological Approach to Visual Perception* (Boston: Houghton Mifflin, 1979).
[17] "China: Police DNA Database Threatens Privacy," Human Rights Watch, May 15, 2017, www.hrw.org/news/2017/05/15/china-police-dna-database-threatens-privacy.

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

[18] Nicole Kobie, "The Complicated Truth about China's Social Credit System," *Wired*, January 21, 2019, www.wired.co.uk/article/china-social-credit-system-explained.
[19] Samantha Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked," *Motherboard*, December 11, 2017, motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn.
[20] For a right-by-right discussion of the impact of AI on human rights, see Filippo A. Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim, "Artificial Intelligence & Human Rights: Opportunities & Risks," Berkman Klein Center for Internet & Society at Harvard University, September 25, 2018, cyber.harvard.edu/sites/default/files/2018-09/2018-09_AIHumanRightsSmall.pdf. One topic we have not touched on here but should acknowledge because it is widely discussed is that of algorithmic fairness, which involves the responsible use of big data and machine learning in many domains of life; see, for instance, Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016).
[21] Vernor Vinge, "The Coming Technological Singularity: How to Survive in the Post-Human Era," *The New York Times Archive*, archive.nytimes.com/www.nytimes.com/library/cyber/surf/1120surf-vinge.html.
[22] Murray Shanahan, *The Technological Singularity* (Cambridge, Mass.: MIT Press, 2015), p. 233.
[23] Sarah Knapton, "DeepMind's AlphaZero Now Showing Human-Like Intuition in Historical 'Turning Point' for AI," *Telegraph*, December 6, 2018, www.telegraph.co.uk/science/2018/12/06/deepminds-alphazero-now-showing-human-like-intuition-creativity/.
[24] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (2016), pp. 484–89.
[25] At the core of reinforcement learning is backpropogation and the Monte Carlo Tree Search. "Backpropagation" is shorthand for "the backward propagation of errors." An error is computed at the output and distributed backward throughout the neural network's layers. For more on backpropagation, see the "Backpropogation" page on the DeepAI website: deepai.org/machine-learning-glossary-and-terms/backpropagation. For the Monte Carlo Tree Search, see Martin Müller (2010), "Challenges in Monte Carlo Tree Search," mcts.ai/.
[26] Silver et al., "Mastering the Game of Go."
[27] "AlphaGo Zero — World's Best Go Player," YouTube video, 2:08, posted by SciNews, October 18, 2017, www.youtube.com/watch?v=4Sm922Xp5N4.
[28] Demis Hassabis and David Silver, "AlphaGo Zero: Learning from Scratch," DeepMind, October 18, 2017, deepmind.com/blog/alphago-zero-learning-scratch/; Chris Duckett, "DeepMind AlphaGo Zero Learns on its Own without Meatbag Intervention," ZDNet, October 19, 2017, www.zdnet.com/article/deepmind-alphago-zero-learns-on-its-own-without-meatbag-intervention/.
[29] Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, et al., "De Novo Structure Prediction with Deep-Learning Based Scoring," in *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts)*, December 1–4, 2018, deepmind.com/blog/alphafold/.
[30] Ian Sample, "Google's DeepMind Predicts 3D Shapes of Proteins," *Guardian*, December 2, 2018, www.theguardian.com/science/2018/dec/02/google-deepminds-ai-program-alphafold-predicts-3d-shapes-of-proteins.
[31] Oren Etzioni, quoted in Larry Greenemeier, "AI versus AI: Self-Taught AlphaGo Zero Vanquishes Its Predecessor," *Scientific American*, October 18, 2017, www.scientificamerican.com/article/ai-versus-ai-self-taught-alphago-zero-vanquishes-its-predecessor/; emphasis added.
[32] Razvan Pascanu, Theophane Weber, Peter Battaglia, Yujia Li, Sébastien Recaniere, and David Reichert, "Agents That Imagine and Plan," DeepMind, July 20, 2017, deepmind.com/blog/agents-imagine-and-plan/.
[33] AlphaStar team, "AlphaStar: Mastering the Real-Time Strategy Game Star-Craft II," DeepMind, January 24, 2019, deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.
[34] Kelsey Piper, "StarCraft Is a Deep, Complicated War Strategy Game. Google's AlphaStar AI Crushed It," *Vox*, January 24, 2019, www.vox.com/future-perfect/2019/1/24/18196177/ai-artificial-intelligence-google-deepmind-starcraft-game.
[35] Vinge, "The Coming Technological Singularity."
[36] April Glaser, "Elon Musk Wants to Connect Computers to Your Brain So We Can Keep Up with Robots," Recode, March 27, 2013, www.recode.net/2017/3/27/15079226/elon-musk-computers-technology-brain-ai-artificial-intelligence-neural-lace; see also "We Are Already Cyborgs / Elon Musk / Code Conference 2016," YouTube video, 5:11, posted by Recode, June 2, 2016, www.youtube.com/watch?list=PLKof9YSAshgyPqlKUUYrHfIQaOzFPSL4&v=ZrGPuUQsDjo.

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

[37] Isobel Asher Hamilton, "Elon Musk Believes AI Could Turn Humans into an Endangered Species like the Mountain Gorilla," *Business Insider*, November 26, 2018, www.businessinsider.com/elon-musk-ai-could-turn-humans-into-endangered-species-2018-11.

[38] Kurt Schlosser, "MIT Student Wows '60 Minutes' by Surfing the Internet and Ordering Pizza — with His Mind," GeekWire, April 23, 2018, www.geekwire.com/2018/mit-student-wows-60-minutes-surfing-internet-ordering-pizza-mind/.

[39] For explorations of such a blended world, see Marx Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).

[40] See James H. Moor, "Four Kinds of Ethical Robots" *Philosophy* Now 72 (2009), pp. 12–14. See also James H. Moor and Terrell Ward Bynum, eds., *CyberPhilosophy: The Intersection of Philosophy and Computing* (Oxford: Basil Blackwell, 2002).

[41] Luciano Floridi and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machines* 14, no. 3 (2004), pp. 349–79. For an exploration of artificial morality and the agency of robots, see Catrin Misselhorn, "Artificial Morality. Concepts, Issues and Challenges," *Society* 55, no. 2 (April 2018), pp. 161–9.

[42] For the philosophy of mind behind what is to come, see David Braddon-Mitchell and Frank Jackson, *Philosophy of Mind and Cognition: An Introduction*, 2nd ed. (Malden, Mass.: Blackwell, 2006); Matt Carter, *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence* (Edinburgh: Edinburgh University Press, 2007); and John Heil, *Philosophy of Mind: A Contemporary Introduction*, 3rd ed. (New York: Routledge, 2012).

[43] Thomas Nagel, *Mind & Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False* (New York: Oxford University Press, 2012), p. 128. The emphasis on the differences between the subjective and the objective standpoint permeates Nagel's work, both in his political philosophy and in his philosophy of mind.

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

[45] Daniel C. Dennett, *Consciousness Explained* (Boston: Back Bay Books, 1992); Daniel C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W. W. Norton, 2018), especially ch. 14.

[46] Dennett, From *Bacteria to Bach and Back*, p. 400.

[47] For the functionalist take on the mind, see Heil, *Philosophy of Mind*, ch. 6; for an early formulation of functionalism, see Hilary Putnam, "Minds and Machines," ch. 18 in *Mind, Language, and Reality: Philosophical Papers*, vol. 2 (Cambridge, U.K.: Cambridge University Press, 1975), pp. 362–385; for influential critical discussion, see Ned Block, "Troubles with Functionalism," in Ned Block, ed., *Readings in the Philosophy of Psychology*, vols. 1 (Cambridge, Mass: Harvard University Press, 1980), pp. 268–305.

[48] Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow* (New York: Harper, 2015), Amazon Kindle ed., p. 293.

[49] Zuboff, *Age of Surveillance Capitalism*.

[50] Harari, *Homo Deus*, p. 314.

[51] Elizabeth Glasure, "Artificial Intelligence Is the Next Big Player in Genomics," Biospace, December 11, 2018, www.biospace.com/article/artificial-intelligence-is-the-next-big-player-in-genomics/; Himanshu Goenka, "Bioterrorism and Gene Editing: Can Crispr Tool Be Used as Biological Weapon in War?," *International Business Times*, December 14, 2016, www.ibtimes.com/bioterrorism-gene-editing-can-crispr-tool-be-used-biological-weapon-war-2460102; see also Antonio Regalado, "Top U.S. Intelligence Official Calls Gene Editing a WMD Threat," *MIT Technology Review*, February 9, 2016, www.technologyreview.com/s/600774/top-us-intelligence-official-calls-gene-editing-a-wmd-threat/. The H5N1 flu strain, for example, kills 60 percent of those it infects. Yet, among humans, it is not highly contagious. In 2011, researchers in the United States and Holland altered the H5N1 genome in a way that made its level of contagion high. A strain like this could "change world history if it were ever set free" by triggering a pandemic, "quite possibly with many millions of deaths." Martin Enserink, "Scientists Brace for Media Storm around Controversial Flu Studies," *Science*, November 23, 2011, www.sciencemag.org/news/2011/11/scientists-brace-media-storm-around-controversial-flu-studies.

[52] Stephen Hsu, "Super-Intelligent Humans Are Coming," *Nautilus*, October 16, 2014, nautil.us/issue/18/genius/super_intelligent-humans-are-coming.

[53] Harari, *Homo Deus*, p. 4; emphasis added

[54] Thomas Piketty, *Capital in the Twenty-First Century* (Cambridge, Mass.: Harvard Belknap Press, 2013).

[55] United Nations, "The World Population Prospects: The 2015 Revision, Key Findings and Advance Tables" (working paper ESA/P/WP.241, Department of Economic and Social Affairs, Population Division, United Nations, July 29, 2015), www.un.org/en/development/desa/publications/world-population-prospects-2015-revision.html.

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press

[56] Max Tegmark, speaking in "Do You Trust This Computer?," YouTube video, 1:18:03, posted by Dr. Caleb Cheung, September 5, 2018, www.youtube.com/watch?v=DVprGRt39yg.
[57] Nick Bostrom, "A History of Transhumanist Thought," *Journal of Evolution and Technology* 14, no. 1 (2005), nickbostrom.com/papers/history.pdf.
[58] Harari, *Homo Deus*.
[59] Steven Pinker, *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress* (New York: Penguin, 2017), Amazon Kindle ed., p. 453.
[60] Ellora Thadaney Israni, "When an Algorithm Helps Send You to Prison." *New York Times*, October 26, 2017, www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.
[61] Benjamin I. Page, Jason Seawright, and Mathew J. Lacombe, *Billionaires and Stealth Politics* (Chicago: University of Chicago Press, 2018); Jane Mayer, *Dark Money: The Hidden History of the Billionaires behind the Rise of the Radical Right* (New York: Doubleday Press, 2016).
[62] The following paragraphs draw on Risse, "Human Rights and Artificial Intelligence."
[63] David Hume, *An Enquiry concerning the Principles of Morals*, ed. by J. B. Schneewind (London: Hackett Publishing, 1983).
[64] Immanuel Kant, *Groundwork for the Metaphysics of Morals*, ed. by Mary Gregor and Jens Timmermann, (Cambridge: Cambridge University Press, 2012).
[65] Another way of thinking about the categorical imperative is that it requires us to always act in ways that would pass a generalization test. Certain actions would be rendered impermissible because they would not hold up if everybody were to take them, as, for instance, stealing and lying would not: there would be no property to begin with if everybody stole, and no communication if everybody reserved the right to lie.
[66] Steve Petersen, "Superintelligence as Superethical," in Patrick Lin, Keith Abney, and Ryan Jenkins, eds., *Robot Ethics 2.0* (New York: Oxford University Press, 2017), pp. 332–7; David Chalmers, "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies* 17, nos. 9–10; see also "What Makes People Happy? / Daniel Kahneman," YouTube video, 9:47, from a discussion with Professor Kahneman at the 2017 Asilomar conference, posted by the Future of Life Institute, January 30, 2017, www.youtube.com/watch?v=z1N96In7GUc.
[67] T. M. Scanlon, "What is Morality?" in *The Harvard Sampler: Liberal Education in the Twenty-First Century*, Jennifer M. Shephard, Stephen M. Kosslyn, Evelynn M. Hammonds, eds. (Cambridge, Mass.: Harvard University Press, 2011), pp. 243–266
[68] For speculation on what such mixed societies could be like, see Tegmark, *Life 3.0*, ch. 5.
[69] For the point about Hobbes, see "Prof. Peter Railton — Machine Morality: Building or Learning?," YouTube video, 33:56, posted by the Artificial Intelligence Channel, September 11, 2017, www.youtube.com/watch?v=SsPFgXeaeLI.

Abstract: What are the implications of artificial intelligence (AI) on human rights in the next three decades? Precise answers to this question are made difficult by the rapid rate of innovation in AI research and by the effects of human practices on the adaption of new technologies. Precise answers are also challenged by imprecise usages of the term "AI." There are several types of research that all fall under this general term. We begin by clarifying what we mean by AI. Most of our attention is then focused on the implications of artificial general intelligence (AGI), which entail that an algorithm or group of algorithms will achieve something like superintelligence. While acknowledging that the feasibility of superintelligence is contested, we consider the moral and ethical implications of such a potential development. What do machines owe humans and what do humans owe superintelligent machines?

Keywords: artificial intelligence, machine learning, deep learning, reinforcement learning, superintelligence, artificial general intelligence, ethical impact agents, implicit ethical agents, categorical imperative, human rights

https://doi.org/10.1017/S089267941900011X Published online by Cambridge University Press