HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Human Rights and Artificial Intelligence: An Urgently Needed Agenda

Faculty Research Working Paper Series

Matthias Risse
Harvard Kennedy School

May 2018
RWP18-015

The views expressed in the **HKS Faculty Research Working Paper Series** are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

www.carrcenter.hks.harvard.edu

Statements and views expressed in this paper are solely those of the authors and do not imply endorsement by Harvard University, the Harvard Kennedy School, or the Carr Center for Human Rights Policy.

CARR CENTER FOR HUMAN RIGHTS POLICY

Human Rights and Artificial Intelligence

An Urgently Needed Agenda

Mathias Risse is Professor of Philosophy and Public Policy. 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. He has also worked on questions in ethics, decision theory and 19th century German philosophy, especially Nietzsche (on whose work he regularly teaches a freshman seminar at Harvard). In addition to HKS, he teaches in Harvard College and the Harvard Extension School, and he is affiliated with the Harvard philosophy department. He has also been involved with executive education both at Harvard and in other places in the world. Risse is the author of *On Global Justice* and *Global Political Philosophy,* both published in 2012.

HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

PAPER
MAY 2018

Contents

Introduction

Artificial intelligence generates challenges for human rights. Inviolability of human life is the central idea behind human rights, an underlying implicit assumption being the hierarchical superiority of humankind to other forms of life meriting less protection. These basic assumptions are questioned through the anticipated arrival of entities that are not alive in familiar ways but nonetheless are sentient and intellectually and perhaps eventually morally superior to humans. To be sure, this scenario may never come to pass and in any event lies in a part of the future beyond current grasp. But it is urgent to get this matter on the agenda. Threats posed by technology to other areas of human rights are already with us. My goal here is to survey these challenges in a way that distinguishes short-, medium-term and long-term perspectives.[1]

[1] For introductory discussions of AI, see Frankish and Ramsey, *The Cambridge Handbook of Artificial Intelligence*; Kaplan, *Artificial Intelligence*; Boden, *AI*. For background on philosophy of technology much beyond what will be discussed here, see Kaplan, *Readings in the Philosophy of Technology*; Scharff and Dusek, *Philosophy of Technology*; Ihde, *Philosophy of Technology*; Verbeek, *What Things Do*. See also Jasanoff, *The Ethics of Invention*. Specifically on philosophy and artificial intelligence, see Carter, *Minds and Computers*. For an early discussion of how the relationship between humans and machines may evolve, see Wiener, *The Human Use Of Human Beings*. That book was originally published in 1950.

AI and Human Rights

AI is increasingly present in our lives, reflecting a growing tendency to turn for advice, or turn over decisions altogether, to algorithms. By “intelligence”, I mean the ability to make predictions about the future and solve complex tasks. “Artificial” intelligence, AI, is such ability demonstrated by machines, in smart phones, tablets, laptops, drones, self-operating vehicles or robots that might take on tasks ranging from household support, companionship of sorts, even sexual companionship, to policing and warfare.

Algorithms can do anything that can be coded, as long as they have access to data they need, at the required speed, and are put into a design frame that allows for execution of the tasks thus determined. In all these domains, progress has been enormous. The effectiveness of algorithms is increasingly enhanced through “Big Data:” availability of an enormous amount of data on all human activity and other processes in the world which allow a particular type of AI known as “machine learning” to draw inferences about what happens next by detecting patterns. Algorithms do better than humans wherever tested, even though human biases are perpetuated in them: any system designed by humans reflects human bias, and algorithms rely on data capturing the past, thus automating the status quo if we fail to prevent them. [2] But algorithms are noise-free: unlike human subjects, they arrive at the same decision on the same problem when presented with it twice.[3]

[2] See this 2017 talk by Daniel Kahneman: https://www.youtube.com/watch?v=z1N96In7GUc On this subject, see also Julia Angwin, “Machine Bias.” On fairness in machine learning, also see Binns, “Fairness in Machine Learning: Lessons from Political Philosophy”; Mittelstadt et al., “The Ethics of Algorithms”; Osoba and Welser, *An Intelligence in Our Image*.

[3] On Big Data, see Mayer-Schönberger and Cukier, *Big Data*. On machine learning, see Domingos, *The Master Algorithm*. On how algorithms can be used in unfair, greedy and otherwise perverse ways, see O’Neil, *Weapons of Math Destruction*. That algorithms can do a lot of good is of course also behind much of the potential that social science has for improving the lives of individuals and societies, see e.g., Trout, *The Empathy Gap*.

For philosophers what is striking is how in the context of AI many philosophical debates reemerge that to many seemed so disconnected from reality. Take the *trolley problem*, which teases out intuitions about deontological vs. consequentialist morality by confronting individuals with choices involving a runaway trolley that might kill various numbers of people depending on what these individuals do. These decisions not only determine who dies, but also whether some who would otherwise be unaffected are instrumentalized to save others. Many a college teacher deployed these cases only to find students questioning their relevance since in real life choices would never be this stylized. But once we need to program self-driving vehicles (which just created their first roadside fatality), there is a new public relevance and urgency to these matters.

Also, philosophers have long puzzled about the nature of the mind. One question is if there is more to the mind than the brain. Whatever else it is, the brain is *also* a complex algorithm. But is the brain fully described thereby, or does that omit what makes us distinct, namely, *consciousness*? Consciousness is the qualitative experience of being somebody or something, its "what-it-is-like-to-be-*that*"-ness, as one might say. If there is nothing more to the mind than the brain, then algorithms in the era of Big Data will outdo us soon at almost everything we do: they make ever more accurate predictions about what book we enjoy or where to vacation next; drive cars more safely than we do; make predictions about health before our brains sound alarms; offer solid advice on what jobs to accept, where to live, what kind of pet to adopt, if it is sensible for us to be parents and whether it is wise to stay with the person we are currently with – based on a myriad of data from people *relevantly like us*. Internet advertisement catering towards our preferences by assessing what we have ordered or clicked on before is a mere shadow of what is to come.

If the mind just is a complex algorithm, then we may eventually have little choice but to grant the same moral status to certain machines that humans have. Questions about the moral status of *animals* arise because of the many continuities between humans and other species: the less we can see them as different from us in terms of morally relevant properties, the more we must treat them as fellow travelers in a shared life, as done for instance in Sue Donaldson and Will Kymlicka's *Zoopolis*.[4] Such reasoning eventually carries over to machines. We should not be distracted by the fact that, as of now, machines have turn-off switches. Future machines might be composed and networked

[4] Donaldson and Kymlicka, *Zoopolis*.

in ways that no longer permit easy switch-off. More importantly, they might display emotions and behavior to express attachment: they might even worry about being turned off, and be anxious to do something about it. Or future machines might be cyborgs, partly composed of organic parts, while humans are modified with non-organic parts for enhancement. Distinctions between humans and non-humans might erode. Ideas about personhood might alter once it becomes possible to upload and store a digitalized brain on a computer, much as nowadays we can store human embryos.

Even before that happens, new generations will grow up with machines in new ways. We may have no qualms about smashing laptops when they no longer perform well. But if we grow up with a robot nanny whose machine-learning capacities enable it to attend to us in ways far beyond what parents do, we would have different attitudes towards robots. Already in 2007, a US colonel called off a robotic land-mine-sweeping exercise because he considered the operation inhumane after a robot kept crawling along losing legs one at a time. [5] Science fiction shows like *Westworld* or *The Good Place* anticipate what it would be like to be surrounded by machines we can only recognize as such by cutting them open. A humanoid robot named Sophia with capabilities to participate in interviews, developed by Hanson Robotics, became a Saudi citizen in October 2017. Later Sophia was named UNDP's first-ever Innovation Champion, the first non-human with a UN title.[6] The future might remember these as historic moments. The pet world is not far behind. Jeff Bezos recently adopted a dog called SpotMini, a versatile robotic pet capable of opening doors, picking himself up and even loading the dishwasher. And SpotMini never needs to go outside if Bezos would rather shop on Amazon or enjoy presidential tweets.

If there indeed *is* more to the mind than the brain, dealing with AI including humanoid robots would be easier. Consciousness, or perhaps accompanying possession of a conscience, might then set us apart. It is a genuinely open question how to make sense of qualitative experience and thus of consciousness. But even though considerations about consciousness might contradict the view that AI systems are moral agents, they will not make it impossible for such systems to be legal actors and as such own property, commit crimes and be accountable in legally enforceable ways. After all, we have a history of treating *corporations* in such ways, which also do not have consciousness.

[5] Wallach and Allen, *Moral Machines,* 55.

[6] https://en.wikipedia.org/wiki/Sophia_(robot)

Much as there are enormous difficulties separating the responsibility of corporations from that of humans involved with them, similar issues will arise with regard to intelligent machines.

The Morality of Pure Intelligence

One other long-standing philosophical problem that obtains fresh relevance here is the connection between rationality and morality. This question emerges when we wonder about the morality of pure intelligence. The term "singularity" refers to the moment when machines surpass humans in intelligence. Since then humans have succeeded in creating something smarter than themselves, this new type of brain may well produce something smarter *than itself*, and on it goes, possibly at great speed. There will be limits to how long this can continue. But since computational powers have increased rapidly over the decades, the limits to what a superintelligence can do are beyond what we can fathom now. Singularity and superintelligence greatly exercise some participants in the AI debate whereas others dismiss them as irrelevant compared to more pressing concerns. Indeed, there might never be a singularity, or it might be decades or hundreds of years off. Still, the exponential technological advancement of the last decades puts these topics on our agenda.[7]

What philosophers think of then is the dispute between David Hume and Immanuel Kant about whether rationality fixes our values. Hume famously thought reason did nothing to fix values: a being endowed with reason, rationality or intelligence (let us assume these are all relevantly similar) might have any goals, as well as any range of attitudes, especially towards human beings. If so, a superintelligence – or any AI for that matter, but the issue is especially troublesome for a superintelligence – could have just about any type of value commitment, including ones that would strike us as rather absurd (such as maximizing the number of paperclips in the universe, to mention an example sometimes brought up in the literature). And how would we know that such thoughts are misguided if indeed they are given that such a superintelligence would be by stipulation massively smarter and thus in particular *different* from us?

[7] Chalmers, "The Singularity: A Philosophical Analysis"; Bostrom, *Superintelligence*; Eden et al., *Singularity Hypotheses*.

As opposed to that, there is the Kantian view that derives morality from rationality. Kant's Categorical Imperative asks of all rational beings not ever to 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 deception of other rational beings (which for Kant would always be too much like pure instrumentalization). In a different way of thinking about the Categorical Imperative it requires of 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 did it, 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. 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. But trashing other rational beings in pursuit of our own interests trash *their* capacities to value, which are relevantly the same capacities whose possession we must value in ourselves. If Kant is right, a superintelligence might be a true role-model for ethical behavior. Since we cannot change human nature, and human nature if intensely parochial in its judgements and value commitments, AI might close the gap that opens when humans with their Stone Age, small-group-oriented DNA operate in a global context.[8]

If something like this argument were to work – and there are doubts – we would have nothing to worry about from a superintelligence. Arguably, we would be rational *enough* for this kind of argument to generate protection for humble humans in an era of much smarter machines. But since a host of philosophers who are smart by contemporary standards has argued against the Kantian standpoint, the matter is far from settled. We do not know what these matters would look like from the standpoint of a superintelligence.

Of course, some kind of morality could be in place with superintelligence in charge even if value cannot be derived from rationality alone. There is also the Hobbesian approach of envisaging what would happen to humans aiming for self-preservation and characterized by certain properties in a state of nature without a shared authority.

[8] Petersen, "Superintelligence as Superethical"; Chalmers, "The Singularity: A Philosophical Analysis." See also this 2017 talk by Daniel Kahneman: https://www.youtube.com/watch?v=z1N96In7GUc

Hobbes argues that even though these individuals would not act on shared values just by thinking clear-mindedly, as they would on a Kantian picture, they would quickly experience the nastiness of life without a shared authority. Far from being vile, as individuals they would feel compelled to strike against each other in anticipation. After all, even if they would know themselves to be cooperative and give the other side the benefit of the doubt as well, they could not be sure that other side would give them that same benefit, and might thus feel compelled to strike first given how much is at stake. Unless there is only one superintelligence, or all superintelligences are closely linked anyway, perhaps such reasoning would apply to such machines as well, and they would be subject to some kind of shared authority. Hobbes's state of nature would then describe the original status of superintelligences vis-à-vis each other. Whether such a shared authority would also create benefits for humans is unclear.[9]

Perhaps T. M. Scanlon's ideas about appropriate responses to values would help.[10] The superintelligence might be "moral" in the sense of reacting in appropriate ways towards what it observes all around. Perhaps then we have some chance at getting protection, or even some level of emancipation in a mixed society composed of humans and machines, given that the abilities of the human brain are truly astounding and generate capacities in human beings that arguably should be worthy of respect.[11] But so are also the capacities of animals, which has not normally led humans to react towards them, or towards the environment, in an appropriately respectfully 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, and that will mean the distinctively human life will receive some protection because it is worthy of respect. We cannot know that for sure but we also need not be pessimistic.

[9] For the point about Hobbes, see this 2016n talk by Peter Railton: https://www.youtube.com/watch?v=SsPFgXeaeLI

[10] Scanlon, "What Is Morality?"

[11] For speculation on what such mixed societies could be like, see Tegmark, *Life 3.0*, chapter 5.

Human Rights and the Problem of Value Alignment

All these matters are in a part of the future about which we do not know *when* or even *if* it will ever be upon us. But from a human-rights standpoint these scenarios matter because we would need to get used to sharing the social world we have built over thousands of years with new types of beings. Other creatures have so far never stood in our way for long, and the best they have been able to hope for is some symbiotic arrangements as pets, livestock or zoo displays. All this would explain why we have a UDHR based on ideas about a distinctively human life which seems to merit protection, at the individual level, of a sort we are unwilling to grant other species. On philosophical grounds I myself think it is justified to give special protection to humans that takes the form of individual entitlements, without thereby saying that just about anything can be done to other animals or the environment. But it would all be very different with intelligent machines. We control animals because we can create an environment where they play a subordinate role. But we might be unable to do so with AI. We would then need rules for a world where some intelligent players are machines. They would have to be designed so they respect human rights even though they would be smart and powerful enough to violate them. At the same time they would have to be endowed with proper protection themselves. It is not impossible that, eventually, the UDHR would have to apply to some of them.[12]

There is an urgency to making sure these developments get off to a good start. The pertinent challenge is the *problem of value alignment*, a challenge that arises way before it will ever matter what the morality of pure intelligence is. No matter how precisely AI systems are generated we must try to make sure their values are aligned with ours to render as unlikely as possible any complications from the fact that a superintelligence might have value commitments very different from ours. That the problem of value alignment needs to be tackled now is also implied by the UN Guiding Principles on Business and Human Rights, created to integrate human rights into business decisions. These principles apply to AI. This means addressing questions such

[12] Margaret Boden argues that machines can never be moral and thus responsible agents; she also thinks it is against human dignity to be supplied with life companions or care givers of sorts that are machines. See https://www.youtube.com/watch?v=KVp33Dwe7qA (For impact of technology on human interaction, see also Turkle, *Alone Together*.) Others argue that certain types of AI would have moral rights or deserve other types of moral consideration; for Matthew Liao's and Eric Schwitzgebel's views on this, see see here: https://www.youtube.com/watch?v=X-uFetzOrsg

as "What are the most severe potential impacts?", "Who are the most vulnerable groups?" and "How can we ensure access to remedy?"[13]

In the AI community the problem of value alignment has been recognized at the latest since Isaac Asimov's 1942 short story "Runaround," where he formulates his famous Three Laws of Robotics, which are there quoted as coming from a handbook published in 2058 (sic!): (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given it by human beings except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

However, these laws have long been regarded as too unspecific. Various efforts have been made to replace them, so far without any connection to the UN's Principles on Business and Human Rights or any other part of the human-rights movement. Among other efforts, in 2017 the Future of Life Institute in Cambridge, MA founded around MIT physicist Max Tegmark and Skype co-founder Jaan Tallinn, held a conference on Beneficial AI at the Asilomar conference center in California to come up with principles to guide further development of AI. Of the resulting 23 Asilomar Principles, 13 are listed under the heading of Ethics and Values. Among other issues, these principles insist that wherever AI causes harm, it should be ascertainable why it does, and where an AI system is involved in judicial decision making its reasoning should be verifiable by human auditors. Such principles respond to concerns that AI deploying machine learning might reason at such speed and have access to such a range of data that its decisions are increasingly opaque, making it impossible to spot if its analyses go astray. The principles also insist on value alignment, urging that "highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation" (Principle 10). The ideas explicitly appear in Principle 11 (Human Values) include "human dignity, rights, freedoms, and cultural diversity."[14]

[13] Ruggie, *Just Business*.

[14] https://futureoflife.org/ai-principles/ On value alignment see also https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/

Insisting on human rights presupposes a certain set of philosophical debates has been settled: there are universal values, in the form of rights, and we roughly know which rights there are. As the Asilomar Principles make clear, there are those in the AI community who believe human rights have been established in credible ways. But others are eager to avoid what they perceive as ethical imperialism. They think the problem of value alignment should be solved differently, for instance by teaching AI to absorb input from around the word, in a crowd-sourcing manner. So this is yet another case where a philosophical problem assumes new relevance: our philosophically preferred understanding of meta-ethics must enter to judge if we are comfortable putting human-rights principles into the design of AI, or not.[15]

Human rights also have the advantage that there have been numerous forms of human-rights vernacularization around the world. Global support for these rights is rather substantial. And again, we already have the UN Guiding Principles on Business and Human Rights. But we can be sure China will be among the leading AI producers and have little inclination to solve the value alignment problem in a human-rights minded spirit. That does not have to defeat efforts elsewhere to advance with the human-rights solution to that problem. Perhaps in due course AI systems can exchange thoughts on how best to align with humans. But it would help if humans went about design of AI in a unified manner, advancing the same solution to the value-alignment problem. However, since even human rights continue to have detractors there is little hope that will happen.

What is in any event needed is more interaction among human-rights and AI communities so the future is not created without the human-rights community. (There is no risk it would be created without the AI community.) One important step into this direction is the decision by Amnesty International – the other AI – to make extensive use of artificial-intelligence devices in pursuit of human-rights causes. This initiative was inaugurated by outgoing Secretary General Salil Shetty, the project leader being Sherif Elsayed-Ali. At this stage, Amnesty is piloting use of machine learning in human-rights investigations, and also focuses on the potential for discrimination within use of machine learning, particularly with regard to policing, criminal justice and access to essential economic and social services. Amnesty is also more generally concerned about the impact of automation on society, including the right to work and livelihood. There

[15] On how machines could actually acquire values, see Bostrom, *Superintelligence*, chapters 12-13; Wallach and Allen, *Moral Machines*.

needs to be more such engagement, ideally going both ways, between the human rights movement and the engineers behind this development.

Artificial Stupidity and the Power of Companies

There are more immediate problems than intelligent machines of the future even though those need to be brought on their way properly. The exercise of each human right on the UDHR is affected by technologies, one way or another. Anti-discrimination provisions are threatened if algorithms used in areas ranging from health care to insurance underwriting to parole decisions are racist or sexist because the learning they do draws on sexism or racism. Freedom of speech and expression, and any liberty individuals have to make up their minds, is undermined by the flood of fake news that engulfs us including fabrication of fake videos that could feature just about anybody doing anything, including acts of terrorism that never occurred or were committed by different people.

The more political participation depends on internet and social media, the more they too are threatened by technological advances, ranging from the possibility of deploying ever more sophisticated internet bots participating in online debates to hacking of devices used to count votes or hacking of public administrations or utilities to create disorder. Wherever there is AI there also is AS, *artificial stupidity*. AS could be far worse than the BS we have gotten all too used to: efforts made by adversaries not only to undermine gains made possible by AI but to turn them into their opposite. Russian manipulation in elections is a wake-up call; much worse is likely to come. Judicial rights could be threatened if AI is used without sufficient transparency and possibility for human scrutiny. An AI system has predicted the outcomes of hundreds of cases at the European Court of Human Rights, forecasting verdicts with accuracy of 79%; and once that accuracy gets yet higher it will be tempting to use AI also to reach decisions. Use of AI in court proceedings might help generate access to legal advice to the poor (one of the projects Amnesty pursues, especially in India); but it might also lead to Kafkaesque situations if algorithms give inscrutable advice.[16]

[16] http://www.bbc.com/news/technology-37727387

Any rights to security and privacy are potentially undermined not only through drones or robot soldiers, but also through increasing legibility and traceability of individuals in a world of electronically recorded human activities and presences. The amount of data available about people will likely increase enormously, especially once biometric sensors can monitor human health. (They might check up on us in the shower and submit their data, and this might well be in our best interest because illness becomes diagnosable way before it becomes a problem.) There will be challenges to civil and political rights arising from the sheer existence of these data and from the fact that these data might well be *privately owned*, but not by those whose data they are. Leading companies in the AI sector are more powerful than oil companies ever were, and this is presumably just the beginning of their ascension.

In the past, status in complex societies was determined first by ownership of land and after the Industrial Revolution by ownership of factories. The ensuing highly inegalitarian structures have not worked out well for many. Unequal ownership of data will have detrimental consequences for many people in society as well. If the power of companies such as Alphabet, Apple, Facebook or Tesla is not harnessed for the public good, we might eventually find ourselves in a world dominated by companies, as depicted for instance in Margaret Atwood's novel *Oryx and Crake* or David Foster Wallace's *Infinite Jest*. The Cambridge-Analytica scandal is a wake-up call here, and Mark Zuckerberg's testimony to US senators on April 10, 2018 revealed an astonishing extent of ignorance among senior lawmakers about the workings of internet companies whose business model depends on marketing data. Such ignorance paves the path to power for companies. Or consider a related point: Governments need the private sector to aid in cyber security. The relevant experts are smart, expensive, and many would never work for government. We can only hope that it will be possible to co-opt them given that government is overextended here. If such efforts fail, only companies will provide the highest level of cyber security.

The Great Disconnect: Technology and Inequality

This takes me to my last topic: AI and inequality, and the connection between that topic and human rights. To begin with, we should heed Thomas Piketty's warning that capitalism left to its own devices in times of peace generates ever increasing economic inequality. Those who own the economy benefit from it more than those who just work there. Over time life chances will ever more depend on social status at birth.[17] We also see more and more how those who either produce technology or know how to use technology to magnify impact can command higher and higher wages. AI will only reinforce these tendencies, making it ever easier for leaders across all segments to magnify their impact. That in turn makes producers of AI ever more highly priced providers of technology. More recently, we have learned from Walter Scheidel that, historically, substantial decreases in inequality have only occurred in response to calamities such as epidemics, social breakdowns, natural disasters or war. Otherwise it is hard to muster effective political will for change.[18]

The original Luddites smashed looms in 19th-century England because they worried about jobs. But so far every wave of technological innovation has ended up creating more jobs than it destroyed. While technological change was not good for everybody, it was good for society as a whole, and for humanity. It is possible that there will be so many jobs that those who develop, supervise or innovatively use technology, as well as creative professions that cannot be displaced, will eventually outnumber those who lose jobs to AI. But clinging to that hope would be naïve because it presupposes a radical overhaul of the educational system to make people competitive. Alternatively, we might hope for some combination of job-creation, shorter working hours so jobs can be shared, but then also higher wages so people can make a decent living. But either way, one can be more hopeful for European countries than for the US, where so many have fallen behind in the race between technology and education and where solidarity at the national level is so poorly entrenched that even universal health care remains contested.[19] How developing countries with comparative advantage in manufacturing and cheap labor will fare in all this is anybody's guess.

Before this background we must worry AI will drive a widening technological wedge into societies that leaves millions excluded, renders them redundant as market participants and thus might well undermine the point of their membership in political community. When wealth was determined by land ownership, the rich needed the rest because the point of land ownership was to charge rent. When wealth was determined by ownership

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

[18] Scheidel, *Great Leveler*.

[19] Goldin and Katz, *The Race Between Education and Technology*.

of factories the owners needed the rest to work the machines and buy stuff. But those on the losing side of the technological divide may no longer be needed at all. In his 1926 short story “The Rich Boy,” F. Scott Fitzgerald famously wrote, “Let me tell you about the very rich. They are different from you and me.” AI might validate that statement in a striking way.

Eventually we might see new Bantustans, as in Apartheid South Africa, or, perhaps more likely, the emergence of separate company-owned entities with wonderful social services from which others are excluded. Perhaps just enough will be given to those others so they do not rebel outright. The fabric of society might dissolve if there are many more people than needed as participants in any sense. Though the world would be rich enough to offer them decent lives, the political will to do so might not be there among the privileged if there are ways of going on that allow the privileged lives without fear of violent disruption. All of that would be seriously bad news from the standpoint of human rights. Scenarios like this are further in the future than the more immediate concerns from the ever-growing presence of algorithms in human life, but probably not as far in the future as the arrival of a superintelligence. Chances are challenges coming from increasing inequality arrive within the next 70 years of the UDHR.

The US is the hub of global technology, including AI, but it indeed has much less practice than, say, many European nations in nation-wide solidarity to help with sustained efforts to make AI beneficial to the whole population. The US has appallingly low social mobility. Studies find that up to 50% of all jobs are now susceptible to automation, including traditionally safe professions such as law, accountancy and medicine. [20] Or as Philip Alston, UN Special Rapporteur on Extreme Poverty and Human Rights, noted about a 2017 official visit to the US:

> Automation and robotization are already throwing many middle-aged workers out of jobs in which they once believed themselves to be secure. In the economy of the twenty-first century, only a tiny percentage of the population is immune from the possibility that they could fall into poverty as a result of bad breaks beyond their own control.[21]

[20] https://rightsinfo.org/rise-artificial-intelligence-threat-human-rights/

[21] http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=22533&LangID=E

We often hear that we should progress with technological change only if it can be shared widely.[22] But as just noted, radical measures against inequality only happen at deeply troubled times, times we would not otherwise wish to live in. The increases in inequality in recent decades, as well as the election of a man who personifies greed, vindictiveness and utter lack of normal empathy do not bode well for any efforts at spreading the wealth in the US, regardless of how nice that sounds at conferences and political events.

We should worry about these increases of inequality also for their impact on human rights. It is hard to overstate what is at stake. Marx was right when, in *On the Jewish Question*, he pointed out that emancipation conceived fully in terms of rights was unappealing. A society built around rights-based ideals misses out on too much. Over the last 70 years the human-rights-movement has often failed to emphasize that larger topic of which human rights must be part: *distributive justice, domestic and global*. AI might eventually jeopardize the very legacy of the Enlightenment because individuality as such is increasingly under siege in an era of Big Data and machine learning. It might also do so since what is threatened here as well is the kind of concern with society as a whole captured in modern thinking about distributive or social justice that became possible only with the spirt of the Enlightenment and technological possibilities opened up by industrialization. I wish I could end on a more uplifting note, and I do not actually think it is "too late." But chances are increasing inequality in combination with AI will be the bane of the next 70 years in the life of the UDHR. Unless, perhaps, enough people see these topics as included in the fierce urgency of now.

On the technological divide, see also https://www.politico.com/agenda/story/2018/02/07/technology-interview-mit-david-autor-000629 And see also http://harvardpolitics.com/world/automation/ On AI and the future of work, also see Brynjolfsson and McAfee, *The Second Machine Age*; Kaplan, *Humans Need Not Apply*.

[22] For instance, at this event: http://futureofwork.mit.edu/

Literature

Binns, Reuben. “Fairness in Machine Learning: Lessons from Political Philosophy.” *Proceedings of Machine Learning Research* 81 (2018): 1–11.

Boden, Margaret A. *AI: Its Nature and Future*. 1 edition. Oxford, United Kingdom: Oxford University Press, 2016.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Reprint edition. Oxford, United Kingdom ; New York, NY: Oxford University Press, 2016.

Brynjolfsson, Erik, and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. 1 edition. New York London: W. W. Norton & Company, 2016.

Carter, Matt. *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence*. 1 edition. Edinburgh: Edinburgh University Press, 2007.

Chalmers, David J. “The Singularity: A Philosophical Analysis.” *Journal of Consciousness Studies* 17, no. 9–10 (2010): 7–65.

Domingos, Pedro. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Reprint edition. Basic Books, 2018.

Donaldson, Sue, and Will Kymlicka. *Zoopolis: A Political Theory of Animal Rights*. 1 edition. Oxford ; New York: Oxford University Press, 2013.

Eden, Amnon H., James H. Moor, Johnny H. Soraker, and Eric Steinhart, eds. *Singularity Hypotheses: A Scientific and Philosophical Assessment*. 2012 edition. New York: Springer, 2013.

Frankish, Keith, and William M. Ramsey, eds. *The Cambridge Handbook of Artificial Intelligence*. Cambridge, UK: Cambridge University Press, 2014.

Goldin, Claudia, and Lawrence Katz. *The Race Between Education and Technology*. Cambridge, Mass.: Belknap, 2008.

Ihde, Don. *Philosophy of Technology: An Introduction*. 1st edition. New York: Paragon House, 1998.

Jasanoff, Sheila. *The Ethics of Invention: Technology and the Human Future*. New York: W. W. Norton & Company, 2016.

Julia Angwin, Jeff Larson. “Machine Bias.” Text/html. ProPublica, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Kaplan, David M., ed. *Readings in the Philosophy of Technology*. 2 edition. Lanham: Rowman & Littlefield Publishers, 2009.

Kaplan, Jerry. *Artificial Intelligence: What Everyone Needs to Know*. 1 edition. New York, NY, United States of America: Oxford University Press, 2016.

———. *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*. Reprint edition. New Haven: Yale University Press, 2016.

Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books, 2014.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3, no. 2 (December 2016): 205395171667967. https://doi.org/10.1177/2053951716679679.

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Reprint edition. New York: Broadway Books, 2017.

Osoba, Osonde A., and William Welser. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, Calif: RAND Corporation, 2017.

Petersen, Steve. “Superintelligence as Superethical.” In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Keith Abney, and Ryan Jenkins, 1 edition., 322–37. New York, NY: Oxford University Press, 2017.

Piketty, Thomas. *Capital in the Twenty-First Century*. Cambridge: Belknap, 2014.

Ruggie, John Gerard. *Just Business: Multinational Corporations and Human Rights*. 1 edition. New York: W. W. Norton & Company, 2013.

Scanlon, T. M. “What Is Morality?” In *The Harvard Sampler: Liberal Education for the Twenty-First Century*, edited by Jennifer M Shephard, Stephen Michael Kosslyn, and Evelynn Maxine Hammonds. Cambridge, Mass., 2011.

Scharff, Robert C., and Val Dusek, eds. *Philosophy of Technology: The Technological Condition: An Anthology*. 2 edition. Malden, MA: Wiley-Blackwell, 2014.

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.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.

Trout, J. D. *The Empathy Gap: Building Bridges to the Good Life and the Good Society*. New York, N.Y: Viking Adult, 2009.

Turkle, Sherry. *Alone Together: Why We Expect More from Technology and Less from Each Other*. Expanded, Revised edition. Basic Books, 2017.

Verbeek, Peter-Paul. *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Illustrated edition edition. University Park, Pa: Penn State University Press, 2005.

Wallach, Wendell, and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. 1 edition. Oxford: Oxford University Press, 2010.

Wiener, Norbert. *The Human Use Of Human Beings: Cybernetics And Society*. Revised edition. New York, N.Y: Da Capo Press, 1988.

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

www.carrcenter.hks.harvard.edu