HUMAN RIGHTS QUARTERLY

Human Rights and Artificial Intelligence: An Urgently Needed Agenda

Mathias Risse

ABSTRACT

The increasing presence of artificial intelligence creates enormous challenges for human rights. Among the short-term challenges are ways in which technology engages just about all rights on the UDHR, as exemplified through use of effectively discriminatory algorithms. Medium-term challenges include changes in the nature of work that could call into question many people's status as participants in society. In the long-term humans may have to live with machines that are intellectually and possibly morally superior, even though this is highly speculative. Artificial intelligence also gives a new relevance to moral debates that used to strike many as arcane.

I. 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 may nonetheless be 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 one's current grasp.

Mathias Risse is the Lucius N. Littauer Professor of Philosophy and Public Administration as well as the Director of the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University. His research addresses many questions of global justice, including human rights, inequality, taxation, trade, immigration, climate change, obligations to future generations, and the future of technology.

Human Rights Quarterly 41 (2019) 1–16

But it is urgent to get this matter on the agenda. Threats posed by technology to other areas of human rights are already very much present. This article surveys these challenges in a way that distinguishes short-, medium-, and long-term perspectives. The main purpose here is to generate more interest in artificial intelligence within the human rights community.[1]

II. AI AND HUMAN RIGHTS

Artificial intelligence (AI) is increasingly present in day-to-day life, reflecting a growing tendency to turn for advice, or turn over decisions altogether, to algorithms. "Intelligence" is the ability to make predictions about the future and solve complex tasks. AI is such an ability demonstrated by machines, in smart phones, tablets, laptops, drones, self-operating vehicles, or robots. Such devices might take on tasks ranging from household support and companionship (including 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:" the availability of an enormous amount of data on all human activity and other processes in the world. Such data allows a particular type of AI known as "machine learning" to draw inferences about what happens next by detecting patterns. Algorithms perform better than humans wherever tested, although 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 unless preventative measures are taken.[2] But algorithms are noise-free:

---

1. For introductory discussions of AI, see The Cambridge Handbook of Artificial Intelligence, (Keith Frankish & William M. Ramsey eds., 2014); Jerry Kaplan, Artificial Intelligence: What Everyone Needs to Know (2016); Margaret A. Boden, AI: Its Nature and Future (2016). For background on philosophy of technology much beyond what will be discussed here, see Readings in the Philosophy of Technology, (David M. Kaplan ed., 2009); Philosophy of Technology: The Technological Condition: An Anthology, (Robert C. Scharff & Val Dusek eds., 2014); Don Ihde, Philosophy of Technology: An Introduction (1998); Peter-Paul Verbeek, What Things Do: Philosophical Reflections on Technology, Agency, and Design (2005). *See also* Sheila Jasanoff, The Ethics of Invention: Technology and the Human Future (2016). Specifically on philosophy and artificial intelligence, see Matt Carter, Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence (2007). For an early discussion of how the relationship between humans and machines may evolve, see Norbert Wiener, The Human Use Of Human Beings: Cybernetics And Society (1988) (1950).
2. See Daniel Kahneman, Discussion at Beneficial AI Conference: What Makes People Happy? (2017), https://www.youtube.com/watch?v=z1N96In7GUc. On this subject, see also Julia Angwin et. al., *Machine Bias*, ProPublica (2016), https://www.propublica.org/

unlike human subjects, they arrive at the same decision on the same problem when presented with it twice.[3]

Philosophical debates have a way of appearing to be disconnected from reality. But in the context of AI, many such debates reemerge with a new kind of urgency. 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 self-driving vehicles (which just caused their first roadside fatality) need to be programmed, 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 fail to recognize what makes humans distinct, namely, *consciousness*? Consciousness is the qualitative experience of being somebody or something, it's "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 soon outdo humans at almost everything: they will make ever more accurate predictions about what book one would enjoy or where to vacation next; drive cars more safely than humans do; make predictions about health before human 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 particular people to be parents and whether it is wise to stay with the person they are currently with—based on a myriad of data from people relevantly like them. Internet advertisement catering towards one's

---

article/machine-bias-risk-assessments-in-criminal-sentencing. On fairness in machine learning, also see Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 Proceedings of Machine Learning Res. 149 (2018); Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 Big Data Soc. 1 (2016); Osonde A. Osoba & William Welser, An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence (2017).

3. On Big Data, see Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think (2014). On machine learning see Pedro Domingos, The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World (2018). On how algorithms can be used in unfair, greedy, and otherwise perverse ways, see Cathy O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2017). 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., J. D. Trout, The Empathy Gap: Building Bridges to the Good Life and the Good Society (2009).

preferences by assessing what one has ordered or clicked on before merely foreshadows what is to come.

If the mind is just a complex algorithm, then there may eventually be little choice but to grant certain machines the same moral status that humans have. Questions about the moral status of *animals* arise because of the many continuities between humans and other species: the less one can see them as different from humans in terms of morally relevant properties, the more they must be treated as fellow travelers in a shared life, as done for instance in Sue Donaldson and Will Kymlicka's *Zoopolis*.[4] Such reasoning may eventually carry over to machines. One should not be distracted by the fact that, as of now, machines have turn-off switches. To survey some possibilities, future machines might be composed and networked 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 well erode. Ideas about personhood might alter once it becomes possible to upload and store a digitalized brain on a computer, much as nowadays human embryos can be stored.[5]

Even before that happens, new generations will grow up with machines in new ways. The typical computer user nowadays may have no qualms about smashing her laptop when it no longer performs well. But people who grow up with a robot nanny whose machine-learning capacities enable it to attend to them in ways far beyond what parents typically do may 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.[6] Science fiction shows like *Westworld* or *The Good Place* anticipate what it would be like to be surrounded by machines one 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 United Nations Development Programme's (UNDP) first-ever Innovation Champion, the first non-human with a UN title.[7] The future might remember these as historic moments. The pet world is not far behind. Amazon founder and CEO Jeff Bezos recently

4. Sue Donaldson & Will Kymlicka, Zoopolis: A Political Theory of Animal Rights (2013).
5. For exploration of these possibilities, see Yuval Noah Harari, Homo Deus: A Brief History of Tomorrow (2017).
6. Wendell Wallach & Colin Allen, Moral Machines: Teaching Robots Right from Wrong 55 (2010).
7. *See Sophia (Robot)*, Wikipedia, https://en.wikipedia.org/wiki/Sophia_(robot).

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.

If there indeed is more to the mind than the brain, dealing with AI, including in the form of humanoid robots, would be easier. Consciousness, or perhaps the possession of a brain *and* a conscience, might then set humans apart. It is a genuinely open question how to make sense of qualitative experiences, 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, there is a long history of treating corporations, which also lack consciousness, in such ways. Much as there are enormous difficulties separating the responsibility of corporations from that of humans involved with them, chances are similar issues will arise with regard to intelligent machines.

III. 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 one wonders about the morality of pure intelligence. The term "singularity" is commonly taken to refer to the moment when machines surpass humans in intelligence.[8] Since then humans will 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 superintelligence can do are beyond what one can fathom now. Singularity and superintelligence are greatly emphasized by 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 (singularity and the moral consequences arising from the existence of a superintelligence) on the human rights agenda.[9]

---

8. For one author who thinks the singularity is near, see Ray Kurzweil, The Singularity is Near: When Humans Transcend Biology (2005).
9. David J. Chalmers, *The Singularity: A Philosophical Analysis*, 17 J. of Consciousness Stud. 7 (2010); Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (2016); Singularity Hypotheses: A Scientific and Philosophical Assessment (Amnon H. Eden et al. eds., 2012).

In order to address the potential moral consequences of AI, philosophers may well think of the dispute between David Hume and Immanuel Kant about whether rationality fixes values. Hume famously thought reason did nothing to fix values: a being endowed with reason, rationality or intelligence (supposing these are all relevantly similar) might have any goals, as well as any range of attitudes especially towards human beings.[10] 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 humans as rather absurd (such as maximizing the number of paperclips in the universe, to mention an example sometimes brought up in the literature).[11] And how would one know that such thoughts are misguided, if it is indeed stipulated that such a superintelligence would be massively smarter, and thus may prioritize different value commitments as compared to humans?

As opposed to that, there is the Kantian view that derives morality from rationality. Kant's "Categorical Imperative" asks of all rational beings never 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).[12] A different way of thinking about the Categorical Imperative requires each person 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 them. For instance, stealing and lying would not be generalizable, and therefore not permissible: 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 that by valuing anything at all one is committed to valuing one's capacity to value. The idea is that things in the world around us are not in some independently given manner valuable, as they could be if we knew that there existed a God who makes them so or if we had reason to think things are valuable by nature much in the same way in which laws of physics apply to them. But these options are not available according to what Kant explores as the limitations of human reason. So that

10. David Hume, An Enquiry Concerning the Principles of Morals (J.B. Schneewind ed., 1983) (1751).
11. First apparently in Nick Bostrom, *Ethical Issues in Advanced Artificial Intelligence, in* Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence (George Eric Lasker, Wendell Wallach, Iva Smit, eds., 2003).
12. Immanuel Kant, Groundwork for the Metaphysics of Morals (Arnulf Zweig trans., Thomas E. Hill, Jr. & Arnulf Zweig eds., (2002) (1785).

leaves human reason itself as the sole source of any kind of value. But if so, then we must appreciate in ourselves that very capacity to value. Therefore, then, trashing other rational beings in pursuit of one's own interests in turn trashes *their* capacities to value, which are relevantly the same capacities whose possession one must value in oneself. For that reason, certain ways of mistreating others lead an actor into a contradiction with herself, in much the same way flaws in mathematical reasoning do. If Kant is right, a superintelligence might be a true role-model for ethical behavior. Since human nature is 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 try to operate in a global context.[13]

If something like this argument were to work—and there are doubts—there would be no reason to worry about a superintelligence. Arguably humans 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 have argued against the Kantian standpoint, the matter is far from settled. Human reason is incapable of imagining 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.[14] Hobbes argues that these individuals would not act on shared values just by thinking clear-mindedly, as they would on a Kantian picture, and that 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 of future wrongs. 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

13. Steve Petersen, *Superintelligence as Superethical*, *in* Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence 322 (Patrick Lin, Keith Abney, & Ryan Jenkins eds., 2017); Chalmers, *supra* note 9. *See also* Kahneman, *supra* note 2.
14. Thomas Hobbes, Leviathan (1651).

each other. Whether such a shared authority would also create benefits for humans is unclear.[15]

Perhaps T. M. Scanlon's ideas about appropriate responses to values would help.[16] The superintelligence might be "moral" in the sense of reacting in appropriate ways towards what it observes all around. Perhaps then humans 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.[17] 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 respectful way. Instead of displaying something like an enlightened anthropocentrism, humans have too often instrumentalized nature. Hopefully a superintelligence would simply outperform human reason in such matters, and that will mean the distinctively human life will receive some protection because it is worthy of respect. There is no way to know that for sure, but there is also no reason to be overly pessimistic.

IV. HUMAN RIGHTS AND THE PROBLEM OF VALUE ALIGNMENT

All these matters are in a part of the future that one cannot know when or even if it will ever materialize. But from a human rights standpoint these scenarios matter because humans would need to get used to sharing the social world they have built over thousands of years with new types of beings. Other creatures have so far never stood in humanity's 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 there is a Universal Declaration of Human Rights (UDHR) based on ideas about a distinctively human life which seems to merit protection, at the individual level, of a sort that humans are unwilling to grant other species. On philosophical grounds it is arguably justifiable 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. Humans control animals because humans can create an environment where animals play a

15. For the point about Hobbes, see Peter Railton, Talk at New York University: Machine Morality (2017), https://www.youtube.com/watch?v=SsPFgXeaeLI.
16. T. M. Scanlon, *What is Morality?, in* The Harvard Sampler: Liberal Education for the Twenty-First Century (Jennifer M Shephard, Stephen Michael Kosslyn, & Evelynn Maxine Hammonds eds., 2011).
17. For speculation on what such mixed societies could be like, see Max Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence 161 (2017).

subordinate role. But this might not be possible with AI. Rules would then be needed for a world where some intelligent players are machines. These intelligent players 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.[18]

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 it is important to try to make sure their values are aligned with human values in order to render as unlikely as possible any complications stemming 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, which was created to integrate human rights into business decisions. These principles apply to AI. This means addressing questions such as "What are the most severe potential impacts?," "Who are the most vulnerable groups?," and "How can one ensure access to remedy?"[19]

The AI community recognized the problem of value alignment as early as 1942 with Isaac Asimov's short story "Runaround," where he formulates his famous "Three Laws of Robotics," which within the story are quoted as coming from a handbook published in 2058:

> (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.[20]

However, these laws have long been regarded as too unspecific, and various efforts have been made to replace them, albeit without any connec-

---

18. 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 Margaret Boden, Panel at Beneficial AI Conference: AI and Ethics (2017), https://www.youtube.com/watch?v=KVp33Dwe7qA. For impact of technology on human interaction, see also Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other (2017). 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 S. Matthew Liao, Presentation, New York University Ethics and Artificial Intelligence Conference: Intelligence and Moral Status (2017), https://www.youtube.com/watch?v=X-uFetzOrsg.
19. John Gerard Ruggie, Just Business: Multinational Corporations and Human Rights (2013).
20. Isaac Asimov, I, Robot, "Three Laws of Robotics" (1970) (1942), https://www.ttu.ee/public/m/mart-murdvee/Techno-Psy/Isaac_Asimov_-_I_Robot.pdf.

tion to the United Nations Principles on Business and Human Rights or any other part of the human rights movement. For example, in 2017 the Future of Life Institute in Cambridge, MA founded by MIT physicist Max Tegmark and Skype co-founder Jaan Tallinn, held a conference on Beneficial AI at the Asilomar conference center in California in order to come up with principles to guide further development of AI. Of the resulting twenty-three Asilomar Principles, thirteen 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). These human values appear explicitly in Principle 11 and include "human dignity, rights, freedoms, and cultural diversity."[21]

Insisting on human rights presupposes a certain set of philosophical debates have been settled: there are universal values, in the form of rights, and it is roughly known 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: one's philosophically preferred understanding of meta-ethics must play a role in deciding whether or not one is comfortable putting human rights principles into the design of AI.[22]

Human rights also have the advantage in that there have been numerous forms of human rights vernacularization around the world, and global support for these rights is rather substantial. Again, the UN Guiding Principles on Business and Human Rights are already in place, but chances are China will be among the leading AI producers and have little inclination to solve the value alignment problem in a human rights spirit. However, that does not have to defeat efforts elsewhere to advance with the human rights

21. *Asilomar AI Principles*, Future of Life Inst. (2017), https://futureoflife.org/ai-principles/. On value alignment see also Ariel Conn, *How Do We Align Artificial Intelligence with Human Values?*, Future of Life Inst. (2017), https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/.

22. On how machines could actually acquire values, see Bostrom, *supra* note 9, 185–227; Wallach & Allen, *supra* note 6.

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 needed, in any event, is more interaction among human rights and AI communities so the future is not created without input from the human rights community. (There is clearly no risk it would be created without the AI community.) One important step in 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. At this stage, Amnesty is piloting the use of machine learning in human rights investigations, and is also focusing on the potential for discrimination within the use of machine learning, particularly with regard to policing, criminal justice, and access to essential economic and social services. More generally, Amnesty is concerned about the impact of automation on society, including the right to work and livelihood. There needs to be more of such engagement, ideally going both ways, between the human rights movement and the engineers behind this development.

V. ARTIFICIAL STUPIDITY AND THE POWER OF COMPANIES

For now there are more immediate problems than intelligent machines of the future. The exercise of each human right on the UDHR is affected by technologies, one way or another. For example, 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 a flood of fake news 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. AI is involved both in the creation and dissemination of such fake-news products.

The more that 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*: 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 percent;[23] and as accuracy increases 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 impenetrable advice whose bases are beyond ready (or perhaps any) human scrutiny. Algorithmic fairness has for good reason started to attract a fair amount of attention.[24]

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 people in the shower and submit their data, and this might well be in one's best interest because some illness becomes diagnosable long before it becomes a problem.) There will be challenges to civil and political rights arising from the sheer existence of such data and from the fact that these data might well be *privately owned,* not by those whose data they are, but by entities other than the ones who generated the data in the first place. Today's 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 (the parent company of Google and its subsidiaries), Apple, Facebook, or Tesla is not harnessed for the public good, humanity might eventually find itself 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 10 April 2018 revealed the astonishing extent of ignorance among senior lawmakers about the workings of internet companies whose business models depend on marketing data. Such ignorance paves the path to power for companies. Consider a related point: governments need the private sector to aid in cyber security. The relevant experts

23. *Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, Vasileios Lampos, Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective* (2016); https://peerj.com/articles/cs-93/.
24. Jane Wakefield, *AI Predicts Outcome of Human Rights Cases*, BBC News, 23 Oct. 2016, http://www.bbc.com/news/technology-37727387.

are smart, expensive, and many would never work for government. One can only hope that it will be possible to co-opt them in service of government security given that governments are overextended here. If such efforts fail, only companies will provide the highest level of cyber security.

VI. THE GREAT DISCONNECT: TECHNOLOGY AND INEQUALITY

The last topic to be discussed here is AI and inequality, and the connection to human rights. The UDHR turns seventy in 2018. That is also a good time to reflect on how there have been numerous instances where technology created the potential for, or inadvertently succeeded in creating, inequality in society, with ensuing implications for human rights. To begin with, there is 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 within it. As such, over time opportunities in one's life will depend ever more on social status at birth.[25] It is also becoming increasingly clear 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, Walter Scheidel has shown 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.[26]

The original Luddites smashed looms in nineteenth-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, one might hope for some combination of job creation, shorter working hours so jobs can be shared, and higher wages so people can make a decent living. Either way, one can be more hopeful for European countries than for the US, where so many have fallen

25. Thomas Piketty, Capital in the Twenty-first Century (2014).
26. Walter Scheidel, Great Leveler: Violence and the History of Inequality from the Stone Age to the Twenty-First Century (2017).

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.[27] How developing countries with comparative advantage in manufacturing and cheap labor will fare in all this is anybody's guess.

Against this backdrop, there is reason to worry that AI will drive a widening technological wedge into societies, excluding millions and rendering them redundant as market participants, thus potentially undermining their membership in political community. When wealth was determined by land ownership the rich needed the masses because the point of land ownership was to charge rent. When wealth was determined by ownership of factories the owners needed the masses 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 there might be new Bantustans, as in Apartheid South Africa, or, perhaps more likely, separate company-owned towns 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 people decent lives, the political will to do so might not be there among the privileged if there are ways of going on that allow them to live 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 created by increasing inequality arrive within the next seventy years of the UDHR.

While the US is the hub of global technology, including AI, it has much less practice than, say, many European nations in creating an environment of nationwide solidarity which helps with sustained efforts to make AI beneficial to the whole population. The US has appallingly low social mobility. Studies find that up to fifty percent of all jobs are now susceptible to automation, including traditionally safe professions such as law, accountancy, and medicine.[28] Or as Philip Alston, UN Special Rapporteur on Extreme Poverty and Human Rights, noted about a 2017 official visit to the US:

---

27. Claudia Goldin & Lawrence Katz, The Race Between Education and Technology (2008).
28. Cate Brown, *The Rise of Artificial Intelligence and the Threat to our Human Rights*, Rights Info (2017), https://rightsinfo.org/rise-artificial-intelligence-threat-human-rights/.

> 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.[29]

It is oft said that technological changes should be allowed to progress only if the resulting benefits can be shared widely.[30] But as just noted, radical measures against inequality only happen at deeply troubled times, times one would not otherwise wish to live in. The increases in inequality in recent decades, as well as the election to the presidency 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.

These increases of inequality are also troublesome 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.[31] A society built around rights-based ideals misses out on too much. Over the last seventy years the human rights movement has often failed to emphasize that larger topic of which human rights must be part: distributive justice, both 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 for society as a whole that is captured in modern thinking about distributive or social justice. Such thinking became possible only with the spirt of the Enlightenment and technological possibilities opened up by industrialization.

---

29. United Nations Office of High Commissioner for Human Rights, *The Rise of Artificial Intelligence and the Threat to our Human Rights* (2017), http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=22533&LangID=E. On the technological divide, see also Nancy Scola, *Is Tech Dividing America? How an Engine of Prosperity Split the Haves From the Have-Nots—And how we can fix it*, Politico (2018), https://www.politico.com/agenda/story/2018/02/07/technology-interview-mit-david-autor-000629; *see also* Nicolas Yan, *Automated Inequality*, Harv. Pol. Rev. (2 Oct. 2016), http://harvardpolitics.com/world/automation/. On AI and the future of work, see also Erik Brynjolfsson & Andrew McAfee, The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (2016); Jerry Kaplan, Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence (2016).
30. *See, e.g.*, AI and the Future of Work Congress, http://futureofwork.mit.edu/.
31. Karl Marx, On the Jewish Question (Helen Lederer trans., 1958) (1844).

VII. CONCLUSION

This article has surveyed challenges for human rights that arise from the increasing presence of artificial intelligence in a way that distinguishes short-, medium-, and long-term perspectives. Some of these challenges are already quite present, others need to be on our radar now even though they may not be relevant for a long time, if ever. Chances are it is the increasing inequality in combination with the production of artificial intelligence that will be the bane of the next seventy years in the life of the Universal Declaration of Human Rights. The human rights community has good reason to put artificial intelligence high on its agenda.