Mathias Risse

Introduction to the Symposium on Ethics and Artificial Intelligence

https://doi.org/10.1515/mopp-2022-0025
Published online June 28, 2022

We live in the digital century. Digital computing with electronic devices has only been available for a few decades, but during that time waves of technological innovation have profoundly affected human possibilities. Much progress in the domain of digital computing is now driven by machine learning, a set of methods that analyze the myriads of available data ("big data") for trends and inferences. Unlike conventional programs, machine-learning algorithms draw on available data sources to learn by themselves. Owing to their sophistication and sweeping applications, these techniques are poised to alter our world radically. Typically, they are what efforts at creating artificial intelligence (AI) amount to nowadays.

A distinction between specialized and general AI is commonly made. Regarding *specialized* AI, at the high end one may think of AI mastering chess or Go. More commonly we encounter it in smartphones (Siri, Google Translate, curated newsfeeds, etc.), home devices (Alexa, Google Home, Nest, etc.), personalized customer services, and GPS systems. Specialized AI is used by law enforcement, the military, browser searching, advertising and entertainment (e.g., recommender systems), medical diagnostics, logistics, finance (from assessing credit to flagging transactions), speech recognition, trade bots, and also in music creation or article drafting (e.g., GPT-3's text generator writing posts or code). *General* AI approximates human performance across such fields. Once general AI is smarter than us it could produce something smarter than itself, and so on, perhaps very fast. That moment is called in technology a *singularity*. It would mean an intelligence explosion with possibly grave consequences that might change the course of history more than anything ever has. The possibility that there could be entities that are not alive in any familiar sense but might nonetheless have properties very different from the kinds of things we typically call "machines" also reveals the limits of our philosophical understanding. Could AI be conscious? Could AI be intrinsically valuable even if it is not conscious? We have difficulties answering such questions because we continue to have profound disagreements in fields such as the philosophy of mind and ethics.

Mathias Risse, Harvard Kennedy School, 79 John F. Kennedy Street, Cambridge, MA 02138, USA, E-mail: mathias_risse@hks.harvard.edu

To be sure, the nature and likelihood of a singularity remain intensely disputed, and we are nowhere near anything like it. But "nowhere near" in the first instance means in terms of engineering capacities rather than time. A few breakthroughs could profoundly transform the field, and many experts seem confident that there will be such breakthroughs over the next few decades. Philosophical engagement with these developments is inevitably constrained by their evolving nature. But while an actual intelligence explosion would bring a kind of change that is hard to anticipate, many philosophical questions about AI are either already upon us or will likely become urgent long before such a singularity, and in fact regardless of it. This special issue of *Moral Philosophy and Politics* is devoted to a set of such questions. The five articles in this issue address and advance several topics around AI that have triggered debates among philosophers and researchers in related fields, each of them making a valuable contribution to the burgeoning field of the philosophy of artificial intelligence.

One major concern about machine learning is that ever more sophisticated algorithms are still working with the same human past. That is, the data on which their recommendations and decisions are based reflect the biases and prejudices that have resulted in enormous inequalities in how humans have fared in their respective societies. Sophisticated algorithms working with the record humankind has created might therefore add a perceived rationality and consistency to patterns of behavior that are deeply unjust. Two of the articles in this special issue can be read as making contributions to that debate in a somewhat corrective spirit, though in rather different ways.

The first article is Nir Eisikovits and Dan Feldman's "AI and Phronesis." The authors argue that the perpetuation of prejudices into the future through the use of historical data is in principle a solvable problem. However, what remains a problem, even if we can meet this challenge, is that the use of AI in everyday decision making – from creditworthiness to police force allocation – will gradually undermine our capacity for making judgments on our own. To make their case, they deploy Aristotle's celebrated account of how phronesis (practical rationality) and the moral virtues develop. The gist of that account is that the habitual exercise of practical judgment allows us to hone virtues gradually. But if we deprive ourselves of practice we will grow up and mature without the kind of learning trajectory that makes us increasingly competent practical reasoners. The consequences could be immensely troubling. Turning over decision making to AI might eventually destabilize the very conditions for becoming a good and flourishing person.

Two things are worth noting about the argument Eisikovits and Feldman make. To begin with, their discussion speaks to the long standing opposition between tech-optimists and tech-pessimists. As far as the domain of judgment is concerned, tech-

optimists see that ever more sophisticated algorithms have the potential of making at least some of us better reasoners. We just need to know the difference between the kinds of questions we can turn over to algorithms and the kinds we cannot, which would include the pondering of complex factual and evaluative matters that ultimately requires prudence or wisdom to be sorted out. But the tech-pessimists will think this thought is illusory. More technology to replace core human competences will increasingly impoverish our possibilities for being human in the first place. Secondly, the fact that Eisikovits and Feldman base their discussion on Aristotle also says something important about the philosophical debates about AI. Some questions that arise here are genuinely new, others continue older debates under a new guise. But the classic repertoire of ideas accumulated in the humanities continue to matter deeply because we are still dealing with human affairs.

Jacob Sparks and Athmeya Jayaram's "Rule by Automation: How Automated Decision Systems Promote Freedom and Equality" can also be understood as a contribution to the debate about how AI and big data perpetuate past prejudices. But their point is not that the real problem is something even worse (and something that cannot be fixed). Instead it is that the use of automated systems to avoid the need for human discretion, specifically in government contexts, can help us achieve the ideal of a free and equal society. To the extent that we aim to achieve a pattern of human relations characterized by the absence of domination and hierarchy, automation can help us achieve these goals. Automated systems have neither wills nor intentions, so they cannot impose them on others and in that sense they cannot dominate. Moreover, by not exercising discretion, an automated decision system would improve consistency and predictability. Nobody's fate would depend on a decision maker's mood or the way they allocate their sympathies. By removing human personality from governmental decision-making, automated systems advance the rule of law and the ideal of a free and equal society.

To be sure, Sparks and Jayaram will face criticism from all those who worry about the ways in which past biases and prejudices provide the very materials fed into automated systems. But like Eisikovits and Feldman these authors could say that those challenges could eventually be resolved if the right kind of effort is expended. And while Eisikovits and Feldman would add that the moment we resolve those challenges is when at the latest we should realize what the real challenge is, Sparks and Jayaram would add instead that this is the moment when the true advantages of using automated systems become visible. So whereas Eisikovits and Feldman align themselves with the tech-pessimists, Sparks and Jayaram are in the tech-optimist camp. The danger for a position like theirs is that we might misjudge *from when on* automated systems will have true advantages

over human decision making. After all, there always is the temptation to turn more matters over "to the computer."

Luise Müller's "Domesticating Artificial Intelligence" is neither in the tech-optimist nor in the tech-pessimist camp, but instead asks how we should go about deploying AI in human societies in ways that are safe. One condition to make that happen is that AI agents need to be aligned with our value-laden cooperative human life, a challenge that has come to be known as the "problem of value alignment." One way to solve that problem is to build "moral machines," machines that are "moral" in whatever ways we might be. That is the solution Müller rejects. Instead she proposes that we need an approach to value alignment that takes seriously the categorically different cognitive and moral capabilities between human and AI agents. Rather than building moral machines we should *domesticate* machines.

Domestication is the process of making other species amenable to life with humans, not as equals, but as auxiliary actors. By aligning their behavior with human values, nonhuman animals have been safely integrated into human society. This solution strikes Müller as attractive because, much like nonhuman animals, AI agents lack moral agency; and just like nonhuman animals, we might nevertheless find ways to train them to live and work among us. We should refrain from even aiming to develop AI for contexts where decision making with squarely moral implications is required. And one can easily see why: we would be aiming to achieve something that has a low likelihood of working out well, and in the process might create machines that do genuine damage.

As we think about how we want AI to enter our lives, it is very sensible to usher in comparisons to our treatment of animals. But in light of the fact that AI might eventually play a large role in our lives, rather than take our cues from our history of domesticating animals for ways to solve the value-alignment problem for machines, it might be better if we reconsider our ways of treating animals. At the species-level animals have always lost out to humans: we restrain them, put them to work, eat them or otherwise use them as resources, put them on display in zoos or walk around holding them on leashes. We are arguably not very responsive to their intrinsic value (as we sit around the table while they are on it), and we get away with it because, species to species, we can overpower them. It is possible that eventually we will be unable to overpower AI, and so now might be be a good occasion to reconsider the intrinsic value of non-human entities in ways that are not tarnished by power relations. It would also be prudent to assume that, were AI ever to dominate us, it might well take its cues on how to fill in that dominance relation from how we have done so with creatures inferior *to us*. Of course, it might never come to that, and if indeed it does not, then Müller's solution would have a

lot going for itself. But perhaps some uneasiness remains as to how durable a solution this is, and at what moral price (vis-à-vis other animals) it comes.

The remaining two papers deal with regulation. The starting point for both is the observation that AI development and application are currently regulated in rather different ways around the world. The EU has long turned itself into a regulations empire that makes genuine efforts to champion the rights of consumers and citizens. The United States has done much less in this domain, leaving much regulatory thinking, especially in the technology domain, to the private sector itself. In China government coopts industry efforts at developing and applying AI for its purposes wherever it sees fit, and regulates them accordingly. In addition, transnational entities have penned a number of non-binding ethics guidelines. In light of this bewildering variety of efforts we must ask: how *should* AI be regulated?

Thomas Ferretti's "An Institutionalist Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation" addresses one aspect of this larger question. He explores the cooperation between government and the private sector to address the ethical dimensions of AI. To that end, he deploys an *institutionalist* approach familiar from political philosophy and business ethics, which advocates a "division of moral labor" between government and private sector. The key idea is that governments are often in the best position to create a just society because, at least in principle, they are more legitimate, stable, and efficient than other actors. Accordingly, in many cases, helping governments build adequate regulation should be the ethical priority of all private agents, including businesses.

So, for one thing, it is not in the first instance the task of the private sector to put such regulation in place or to make sure it is. Instead, it is the task of government. At the same time, the private sector can reasonably be expected to comply with government efforts to regulate that domain, and also to encourage and support its efforts. This could mean, for instance, that voluntary standards developed in the private sector become in due course legally binding on everyone. The exception to all this is when government is incapable or unwilling to take on this task, in which case self-regulation is the only option. Under such circumstances, the private sector has a different kind of duty.

Eva Erman and Markus Furendal's "The Global Governance of Artificial Intelligence: Some Normative Concerns" addresses a different part of the question about regulation. The status quo in AI regulation sketched above is unlikely to ensure that global governance of AI will be even minimally democratic and fair. Erman and Furendal do not mean to offer a first-order theory of democratic and fair AI governance, but instead propose a theoretical meta-framework through several desiderata. First, much novel thinking about democracy is needed to make sure AI is regulated through democratic processes in the first place and subsequently

strengthens rather than undermines democratic societies. As Erman and Furendal say, democracy should be theorized in holistic ways, so that we can recognize how different core values interact and are tied together, what the relationship is between authorized and mandated entities, and then also the normative difference between law and policy-making. Secondly, AI regulation should address not only how fairly the technology treats individuals, but also the indirect distribution effects created. Finally, addressing these problems requires not only new institutions to regulate AI but a comprehensive review of political and economic institutions.

This approach to AI regulation is not only intended to improve upon regulatory efforts as they currently unfold in the world, Erman and Furendal are also trying to shift the focus of the debate about regulation. On a more typically taken approach, it is ultimately the hypothetical AI agent itself that is the subject to be governed through regulation. Designing and reforming political institutions is instrumentally important to the extent that it serves the prioritized goal of AI safety. By contrast, Erman and Furendal seek to ground the debate about regulation in the question of how political institutions influence the development and deployment of AI technology, and how they ought to be arranged so as to realize a broad array of goals and ideals. In other words, the key question for these authors is not how to rein in a potential super-intelligent AI system, but how to govern the global actors involved in developing and deploying AI technology in general.

Here's one rather dramatic way of seeing why AI regulation is such an important topic. The *Fermi Paradox*, named after physicist Enrico Fermi, is the apparent contradiction between the plausibly high probabilities for the existence of extraterrestrial life, on the one hand, and the complete absence of credible evidence for such life, on the other. Possibly the constellation of conditions needed for life to emerge is so extraordinary that, the vast size of the universe notwithstanding, either we are alone after all or occurrences of life are so rare that such lack of evidence is to be expected. But according to another resolution of the paradox, this lack of evidence stems from the fact that intelligent life tends to perish after a short time (by cosmic standards). Sometimes this happens accidentally: asteroids might hit, the nearest sun expire, and other such things. But typically perdition comes about as self-destruction in the very exercise of intelligence. Intelligent life, that is, tends to create technology that eventually brings about its own destruction, and does so *before* this intelligent life manages to connect to intelligent life on other planets (which is why we find no evidence of intelligent life elsewhere even though it is likely that there is such life).

The Fermi Paradox should make us pause as we look back at the stunning amount of technological innovation we have witnessed in the last several hundred years, especially throughout the 20th century and then into the 21st. At some level

it would seem a trivial thing to say that what it takes, at a minimum, for technology to be part of a safe human future is for there to be broad societal engagement to make sure technology is developed that actually benefits humanity as a whole. But the realities in the domain of regulation do not reflect this trivial insight.

Among the groups that do take the regulation of technology – and reflection about it at the level of the community as a whole – rather seriously are the Amish in the United States. Descendants from radical sixteenth-century European Protestants and named for their leader Jakob Ammann, the Amish are pacifist Christians who fully separate church and state. Their communities bar technologies they suspect of weakening community ties, strengthening dependence on government or surrounding communities of non-believers, or threatening their pursuit of virtuous lives. The Amish inhabit houses disconnected from the electric grid, cultivate land with horse-drawn machinery and get around in buggies, renounce insurance schemes including social security, do without TV and radio (much less Internet), and limit the use of phones. My point is not to advocate Amish lifestyle. But what the rest of us can learn from them is that it is only for the time being that we get to choose what technologies to develop and use: sooner than one might think the technologies we have deployed will limit what ways of being human are available to us. One way or another, putting regulatory structures in place means setting the stage for how this is sorted out later. So we had better get AI regulation right.

The philosophy of artificial intelligence is now a burgeoning field, and one that goes through the typical dynamics of a relatively new field. Agendas and major topics are still emerging, and what will be important in the long run will to some extent also be determined by technological innovation itself. That said, the themes covered in this special issue will likely remain important. And these five articles themselves make important contributions to advance those themes.