Nir Eisikovits* and Dan Feldman

AI and Phronesis

https://doi.org/10.1515/mopp-2021-0026
Published online August 30, 2021

Abstract: We argue that the growing prevalence of statistical machine learning in everyday decision making – from creditworthiness to police force allocation – effectively replaces many of our humdrum practical judgments and that this will eventually undermine our capacity for making such judgments. We lean on Aristotle's famous account of how *phronesis* and moral virtues develop to make our case. If Aristotle is right that the habitual exercise of practical judgment allows us to incrementally hone virtues, and if AI saves us time by taking over some of those practical judgments, or if its pattern recognition capacities are very good at learning that kind of behavior – we risk innovating ourselves out of moral competence with the introduction of AI.

Keywords: artificial intelligence, phronesis, judgment, Aristotle, algorithmic bias, technological optimism

In this paper, we argue that a key reason to worry about AI is that it undermines our capacity for practical judgment. By gradually taking over some of the contexts in which we exercise what Aristotle called *phronesis*, AI calls into question important aspects of our moral development. Part 1 defines artificial intelligence, briefly surveys the history of the technology and provides detail about the rise of statistical machine learning, the most prevalent variant of AI. Part 2 explores and questions a popular worry about the rise of AI – that it exacerbates existing social and economic inequalities. Part 3 leans on Aristotle's famous "function argument" to suggest that the key reason to worry about AI is that it undermines our capacity for practical judgment. Part 4 claims that this decline in our capacity for judgment is made worse by AI mediated technology, which prompts us to pay greater attention to our own lives and projects.

***Corresponding author: Nir Eisikovits**, Applied Ethics Center, University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA 02125, USA, E-mail: Nir.eisikovits@umb.edu
Dan Feldman, Applied Ethics Center, University of Massachusetts Boston, 100 Morrissey Blvd., Boston, MA 02125, USA, E-mail: djfeldman57@gmail.com

1 Artificial Intelligence and Statistical Machine Learning

Artificial intelligences are machines that demonstrate human-like performance on tasks that humans are generally good at. For example, humans are generally very good at navigating crowded, complex spaces in order to achieve a goal, such as getting from one side of a busy train platform to the other while conversing with a friend. It has generally been impossible to program computers with enough commonsense knowledge about the world and with sufficiently sophisticated perceptual, planning and reasoning mechanisms to enable them to demonstrate human-level performance on day-to-day tasks such as this. The overriding goal of the AI research program has been to achieve human-level performance on just these sorts of tasks. Today, computers are beginning to demonstrate remarkably human-like capabilities but have yet to achieve anything like autonomous, self-conscious, competent functioning in the natural world. Nevertheless, the capabilities that are being commercialized are being called "artificial intelligence" and are hailed in the press and sold in the marketplace as such.

The modern artificial intelligence research program emerged after World War II, enabled by earlier work in cybernetics and the development of electronic digital computers.[1] By the late 1950s, computers were becoming commercially available and computer science – the rigorous analysis of computational techniques – was becoming an academic field. In 1956, a small group of researchers, with funding from the Rockefeller Foundation, convened a conference at Dartmouth College that served to name and to launch AI as a field (see McCarthy et al. 1955). The post-war period also saw the continuation of research and development in cybernetics and the building of practical systems integrating humans and machines to achieve a functional purpose. By the early 1960s, digital computers were central to how these hybrid human-machine systems were being imagined and designed. A cybernetic system, an anti-aircraft gun control system, for instance, would integrate human targeting and firing with machine target tracking and aiming (see Mindell 1995).

Numerous articles in the popular press, pronouncements from respected scientists and public hand wringing by prominent technology company executives and entrepreneurs evince a palpable anxiety about the current trajectory of AI. The late physicist Stephen Hawking and the very active and successful entrepreneur

[1] There is a rich and varied historiography of computing, ranging from surveys of key technical developments (see, e.g., Randell 1982) to detailed investigations into the relationship between the Cold War and the computing community (see, e.g., Edwards 1996).

Elon Musk, to name two, have been vocal in their concerns about the "existential threat" that AI represents (see Domonoske 2017; Love 2014). They are worried about the emergence of machines that are cognitively superior to humans, control the physical world, reproduce and have their own agendas. These functionally superior entities would have little use for humans and would certainly not accept human oversight. In fact, they may have no use for humans at all, leading to speculation about how our species would fare. The human relationship to other species is not an encouraging example of how cognitively superior entities treat others. Will we be the AI's pets? Pests to be exterminated? Irrelevant and so harmed incidentally as the AI pursues its own goals?

Fascinating and perhaps frightening as such speculation is, it is a significant leap from the current capabilities labeled AI to the dystopia of super-intelligence (cf. Bostrom 2016). It may be more useful, at least for now, to engage with the challenges and implications of AI as another example of the complex consequences of the human development of tools.

From this perspective, AI occupies one end of a spectrum of implementation options for achieving useful machines. The spectrum ranges from essentially human with simple mechanical tools (such as hammers and wheels) to fully digital with humans setting goals and leaving implementation to the machines. An artificial general intelligence would be as capable of functioning in the world as a human and, at the limit, of setting its own goals. In fact, autonomous goal setting might be the characteristic that moves such a machine off this spectrum and prompts new anxieties. Certainly off this spectrum is super-intelligence – artificial general intelligence that *exceeds* human capabilities across the broad range of human cognitive abilities, the sort of machine that leads to the existential anxiety discussed above. While one can be concerned about super-intelligence, that is not our concern. Instead, we wish to understand the implications for being human of the emergence (not to say rise) of intelligent machines that approach and often exceed the human capability for recognizing patterns.

The broad AI research program introduced above encompasses a range of cognitive capabilities. While there are certain cognitive tasks that computers have long performed and have done so dramatically more effectively and efficiently than people, such as arithmetic, sampling, or responding to quantifiable changes in the environment, the research program contemplates automated learning, planning and reasoning based on a "commonsense" understanding of the world.[2]

[2] See, for example, Minsky's (1955, 9) research abstract in McCarthy et al. where he suggests that an AI machine "would tend to build up within itself an abstract model of the environment in which it is placed … [It] could first explores solutions within the internal abstract model … and then attempt external experiments … These external experiments would appear to be rather clever."

The learning part of the agenda has made the most visible progress in the last two decades or so. By the mid-1990s, statistical machine learning – an approach to recognizing patterns – started showing dramatic results. In the time since, known but previously impractical computational neural networks extended the power and effectiveness of the pattern recognition toolkit. Statistical machine learning (ML) has come to dominate the popular understanding of AI.

In fact, based on journalistic and commercial activity, we see a two-dimensional telescoping of AI. First, the rich AI research agenda that includes planning, reasoning, learning, knowledge representation and other similarly broad investigations has often been reduced to learning. Second, learning itself has often been reduced to one technique (multiple hidden layer neural networks, called deep learning) in one branch of the field – statistical machine learning. There are exceptions to this generalization: autonomous robots, notably in the form of autonomous motor vehicles, must plan their actions and reason about their environment. Even in these applications, though, deep learning has become a pivotal enabling technique and garnered much attention.[3] Nevertheless, the agenda of the fully developed AI research program would be substantially impoverished if we understood AI to wholly consist of statistical machine learning.

A substantial fraction of deployed statistical machine learning is based on algorithms that emerge when one asks an apparently simple question: given a set of facts that have already been placed into categories, in which category should I place a new fact when I run across it? For instance, if I can describe the leaves of plants and categorize them by their color, shape and size can I then, upon encountering a previously unknown leaf, correctly identify it as being of one type or another? Key concepts then are to select a pattern-recognizing algorithm and construct a pattern recognizer, and then to apply that algorithm to new data. The first is referred to as training and the second as predicting.[4]

Of the many machine learning algorithms, so-called deep learning has come to represent both the potential and the pitfalls of the field. Deep learning algorithms are all based on neural networks, and neural networks, in turn, are based on a still superficial understanding of the massive connectivity between neurons in organic brains. They are models of a biological system and as such they seem to reproduce certain aspects of human learning, such as pattern recognition.

[3] The website iiot-world.com has a brief summary of machine learning applied to the autonomous driving problem ("Machine Learning Algorithms in Autonomous Driving") at https://iiot-world.com/machine-learning/machine-learning-algorithms-in-autonomous-driving/ while Fridman et al. (2018) have written about the substantial challenges in achieving autonomous driving.

[4] Machine learning techniques (algorithms) can be divided into types: clustering, categorization and regression. For a reasonably accessible introduction to the field for the technically inclined, see Witten, Frank, and Hall (2016).

Pattern recognition algorithms based on statistical machine learning have become essential to the administration of numerous governmental and commercial activities. In many cases they perform evaluative functions previously reserved for humans. Examples include teacher evaluations (see Pal and Pal 2013), the assessment of the likelihood of criminal recidivism (see Wang, Mathieu, and Cai 2010), the granting of credit (see Chawla 2018) and detection of fraudulent credit transactions (see Awoyemi, Adetunmbi, and Oluwadare 2017), the identification of tumors in radiographs (see Qin et al. 2018) and whether or not a very sick patient should be referred to hospice care (see Avati et al. 2018). The adoption of machine learning for these tasks relies on the belief that computers are dispassionate and unbiased implementers of objectively appropriate decision-making criteria. However, as substantial current scholarship is beginning to show, these assumptions are not necessarily justified.

The statistical underpinnings of machine learning do not guarantee either objectivity or accuracy. The purpose of training is to construct an algorithm that successfully induces generalizations about certain types of data from a (large) number of specific instances. The training phase of the modeling process depends on large, accurately labeled datasets, the choice of genuinely meaningful criteria, and careful technical steps to ensure that the algorithm works on data that it hasn't been trained on. The model must make usefully accurate predictions about new data and a model maintenance process must incorporate mechanisms for correcting it based on new data – that is, the model must be periodically re-trained.

2 AI and Fairness

Many critiques of the deployment of machine learning focus on methodological failures in the training and predicting stages. However, aside from simply doing the job poorly – using too little training data, selecting less relevant features, testing the model's predictive power inadequately (or not at all!) and failing to retrain based on new data – many machine learning models embed the (often unconscious) biases of the people who construct them.

Biased and inaccurate machine learning can lead to unfairness in the form of reduced opportunity, unjustified denial of various social and economic benefits and skewed distribution of resources. Cathy O'Neil (2016) provides a number of in-depth examples of specific applications of machine learning that she refers to as "weapons of math destruction (WMDs)." WMDs are mathematical models embedded in automated systems and applied in lieu of human judgment in ways that affect human outcomes. These particular models are opaque (not subject to inspection by the subjects they affect), self-reinforcing (they tend to create social outcomes that lead to similar social outcomes in the future), without mechanisms for correction and, crucially, embed the biases of their designers. O'Neil notes that "when [she] considers the sloppy and self-serving ways that companies use data, [she is] often reminded of phrenology" (2016, p. 121). Her concern is that the modeling is "sloppy" and "self-serving" (and hence pseudo-scientific) rather than that it is intrinsically problematic. That is, good models well applied are acceptable; it is poorly designed and implemented models deployed on a large scale that are troubling.

The flaws in ML models can (and, O'Neil and others would argue, often do) contribute to the preservation and perpetuation of social and economic disadvantages by encoding the model developers' biases, reinforcing bias by providing more instances of it and, due to their opacity – are resistant to criticism and correction (see also Danks and London 2017; Eubanks 2018; Noble 2018). Key to these failings is that the models are models because they generally use *proxies* for the phenomenon they purport to measure. They attempt to make a *prediction* about someone's propensity to reoffend, for example, based on socioeconomic facts that are correlated with that person and *others* who have or have not reoffended. Scientists and statisticians are well aware of the danger of confusing correlation with causality and those competent at statistical inference carefully guard against it.

Statistical machine learning, as noted above, is a collection of techniques for recognizing patterns, and deep learning is one of these techniques. It is distinguished from many others by both its ability to recognize patterns that often are not obvious to more typical statistical techniques and by its opacity. It is generally impossible to understand how it is that any particular neural network decides to classify inputs into one category or another. This confounds any disinterested attempt to evaluate the operation of these algorithms. However, even this problem is starting to be addressed (see Kuang 2017). The failures that critics like O'Neil focus on result from poorly designed or implemented modeling. Suppose, however, that statistical techniques, even the most challenging to understand, can be applied rigorously and yield systems that are not intrinsically biased and harmful and that can be inspected in order to understand their operation. Assume, for the sake of argument, that concerns around AI, machine learning and equity can be addressed.[5] Would this exhaust our worries about the technology?

[5] Consider, for example, the intriguing claim made by Kearns and Roth (2019) that some such "fixes" to algorithmic bias can be integrated into the design of the algorithms themselves by quantifying and then incorporating moral restrictions into the code.

3 AI and Phronesis

While AI raises concerns about equity and fairness, as we explained in the previous section, we think that these are not the most significant reasons to worry about the technology. Many of these failures can, in principle, be addressed by revising the algorithms such that they no longer encode the prejudices of those who wrote them, and by putting in place a quality control mechanism that keeps the algorithms in check after they have started working. In fact, this is already happening: face recognition algorithms used to be better at recognizing the faces of white men than at recognizing anyone's else face. When this became known, the ensuing public outcry quickly resulted in the algorithms being tweaked – more images of men and women of color were added to the machine learning process, which duly improved (see Lohr 2018; Sumagaysay 2018). And, it goes without saying, the algorithms would have not had this problem in the first place if there had been more coders who were not white men.

To put it differently, AI frequently reflects and perpetuates the biases of its creators, but these biases can be pointed out and the algorithms can be corrected. And AI, in spite of anxieties prevalent in popular culture, is not a brain. It does not have an amygdala that organically incorporates unconscious bias and anxiety into decision making. There is nothing inherent to an algorithm that inclines it to morally faulty decisions. Thus, its judgments may, in time, actually become fairer than ours.

Our question, then, is this: If algorithmic bias concerns about AI were eliminated, would there be something left to worry about? To put it more sharply, if AI decisions became fairer than typical human decisions, would there be any residual discomfort with the technology? And, if there is, is that discomfort philosophically interesting or just a function of our difficulty to get psychologically accustomed to new modes of decision making?

The most serious concern about the rise of AI can be gleaned with the help of Aristotle's famous "function argument" (1999, Book I). Very broadly, that argument runs something like this: Human actions typically aim towards an end or "good" that is beyond the action itself (we go to the dentist to keep our teeth healthy, we avoid certain foods to keep our hearts healthy and so on). But there must be a "highest good" in the realm of action, something desired for its own sake and not for the sake of something else, otherwise all desire and striving would be futile. We usually call this highest good "happiness". The happiness or *eudaimonia* of X consists in the fulfillment of X's function. A good X, a flourishing X, a happy X is one that does its work (*ergon*) and fulfills its function (a good shoe is one that gets us comfortably from one place to another and good car is one that transports us with a mix of elegance and practicality). Indeed, everything we encounter has a function. Humans must have a function too. The function of human beings is the excellent exercise of rationality. A well-functioning human being is one that uses reason well, in order to modulate wants and desires. That regulation – that exercise – the ability to rationally weigh particulars in different areas and decide how much of a desire we realize and when, is the key to human flourishing. What we are when we are well-functioning humans is creatures who can exercise practical judgment – creatures who can navigate particulars well – and it is by the repeated exercise of this judgment and the navigation of particulars that we gradually become possessed of the moral virtues, that we acquire them as second nature, that we become full-fledged human beings and, with some luck, flourish.

The argument is more complicated than this (for example, Aristotle speaks of two possible exercises of rationality – passively following others' good judgment and actively conceiving of the good oneself – such that one begins with the former and moves to the latter.) But for our purposes we can stick with this sketch.

There are, of course, well known problems with Aristotle's function argument. The assumption that there must be a highest good is questionable (Aristotle says that without a highest good all desire and striving would be futile, but perhaps all desire and striving just are futile or at least they are not made meaningful by a good outside of themselves); it's not clear why humans as such must have a function even if all human professions (carpenters, brick layers, programmers) or all human body parts (the eye, the ear) have one. It's not clear that a creature's happiness must be understood in terms of its function (the early utilitarians and, later, Freud would argue that it should be understood in terms of pleasure and pain or desire satisfaction).[6]

And yet Aristotle's basic insights – that we are judgment making creatures, that making judgements requires weighting particulars, that there is connection

[6] For an excellent discussion see Roochnik (2013), especially chapters 5 and 6. Our argument assumes but does not explicitly defend one position in a long-standing debate in Aristotle scholarship, between intellectualist and inclusionist understandings of *eudaimonia*, or human flourishing. Briefly, the former position, usually traced to Aristotle's argument in Book 10 sections 6–8 of the *Nicomachean Ethics* holds that human flourishing is primarily about, and is indeed defined by, *theoria* or contemplative activity. The contemplative life can involve and perhaps even requires some ethical activities to sustain it, but a person's happiness consists in the realization (or as close as possible) of the life of reflection; ethical virtue will be merely instrumental to that purpose. Inclusionists, on the other hand, argue that there can be more than one good that is sought for its own sake, and that flourishing will consist in a combination of ethical and contemplative virtues. For a leading intellectualist account see Kraut (1989). For examples of the inclusivist position see Roche (1988) and Annas (1999). We proceed from inclusivists assumptions here – that part of what makes a human life valuable and happy is the exercise of phronesis and that the development and practice of moral virtues are good for their own sake.

between making these practical judgments over time and what it means to be good or have good character traits – are powerful. And if we take those insights seriously it follows that reducing the scope in which we can make such judgments is problematic; it destabilizes the very conditions for becoming a good, flourishing person.

Knowing how much I should donate to the Patrolmen's Benevolent Association when they call for a contribution to benefit orphans of police officers requires the capacity to quickly assess whether the information in the request is reliable, how much money I still need to spend on other obligations this month, how much money I have already given to charity, and where this cause stands in relation to other charitable causes. Someone who can quickly weigh all of these factors and give the right amount of money to the right cause for the right reasons and at the right time is generous. Someone who gives nothing is stingy. Someone who bankrupts herself is lavish. The latter two are failures of judgment. Both involve a coarseness in navigating the particulars. Aristotle tells us that it is through the process of making these judgments in the same way a generous person would make them that we ultimately become generous. The same is true about hiring decisions and about punishment decisions and about creditworthiness decisions and decisions made in myriad other contexts.

Who to hire, who to fire, whether to approve a loan, where to allocate police forces, how much to punish, how much to give to charity are all decisions that require the weighting of particulars and, in the end, require virtue to be done well. Our life circumstances are such that many of these judgments are made at work in our professional capacity. But AI is gradually replacing practical decision making in our work lives. As we write this, employment decisions in large chain franchises, loan and credit approval decisions, and charitable giving decisions are being farmed out to algorithms. And the algorithms are doing an increasingly competent job with these decisions. The trend is likely to expand, with more and more aspects of local government (tax assessments, decisions about licensees) becoming automated (see Hughes 2017). Whether or not the algorithms are actually "making judgments" is a quandary we don't have to enter.

They are certainly *replacing* judgments and it's the decline in human capacity we are focused on here rather than the status of algorithms.

Algorithms are eliminating many of the contexts we have for weighting particulars and thus exercising our practical wisdom. Middling management jobs are the opportunities many of us have to hone our ability to make judgments, and many of these jobs are being automated.

The psychologist Barry Schwartz and political scientist Kenneth Sharpe (2011) write of a Pennsylvania judge, Lois Forer, who sentenced a felon to less than the minimum two-year term for armed robbery. She did this because the felony perpetrated was a non-characteristic, anxiety driven slip up. The felon, “Michael”, lost his job, fell into a panic about how he would feed his family and robbed a cab driver (using a toy gun). Forer gave Michael an 11-month sentence and helped keep his family intact by mandating that he work during the day and return to jail at night. This “wise improvisation” by the judge seemed like a successful compromise and an appropriate weighting of the relevant circumstances. Michael completed his sentence and found a new job; he and his family got back on track. But the prosecution appealed the sentence, the higher court mandated the minimum penalty, judge Forer dutifully imposed it, and then she resigned from the bench. She quit because she had become a judge who was not allowed to judge. Being on the bench was meaningful for her, among other things, because it allowed her to use her judgment. The work was no longer of interest once that capacity had been curtailed. Forer was demoralized because, at least for her, judging itself became demoralized.

This case provides a helpful analogy for thinking about the trouble with AI. Sentencing guidelines are not, strictly speaking, machine learning algorithms. But they are close enough. Like algorithms, they are decision making procedures that obviate and replace human judgments. As in the case of sentencing guidelines, AI based decision making is gradually replacing the need to navigate, weigh and assess stories. The algorithms are likely to do a good, fair, maybe even nuanced job. But that ability to navigate, weigh and assess stories is at the heart of what it means to be an active human who realizes her function or, to use Aristotle’s language, “does her work”. When machines do it for us those tasks will become demoralized and we will become demoralized. An algorithmic sentencing decision, teacher termination decision or loan approval may be both fair and successful, yet demoralizing to the official who used to make it. Put differently, AI risks demoralizing us and our activities because a big part of what it means to be moralized is to use practical judgment or phronesis.

If Aristotle is right that habit instills morally excellent behavior – if the habitual exercise of practical judgments allows us to incrementally hone virtues, and AI saves us time by taking over some of those practical judgments, or if its pattern recognition capacities are very good at learning that kind of behavior – we risk innovating ourselves out of moral competence with the introduction of AI. We will, simply, lose the habit. Though Aristotelian ethics are not algorithmic like Kantian or utilitarian accounts, moral capacity and skill à la Aristotle – developed as they are by practice and habit – are algorithmizable. In fact, the habitual weighting of particulars through time with small tweaks based on prior interactions – the essence of virtue acquisition for Aristotle – is also a perfect description of a machine learning model. This does not mean the machine becomes virtuous – it is certainly hard to argue that the conditions for virtue that

Aristotle enumerates (acting knowingly, choosing the action for its own sake, acting from a stable state of character) are met (Aristotle 1999, Book II); it just means that our practical judgment can be replaced by machines. In other words, if the person who has practical wisdom, the *phronimos,* is one who navigates particulars well, one who assigns appropriate weight to them based on context (Aristotle 1999, Book VI),[7] AI can emulate what that person does. In creating an algorithm, weights are assigned to different aspects of the task or situation the machine is confronted with, and the machine gradually adjusts these weights through continuous iterations. Suppose we are talking about locating a medical advisor for a serious disease: practical judgment would weigh expertise, availability, bedside manner etc. and revise judgments about the preferred expert based on the efficacy of the interactions they yield. An algorithm can be written to approximate that process – at least well enough.[8]

Aristotle argues that human flourishing or *eudaimonia* is achieved through work – by practicing the capacities that, like our ability to make practical judgments, make us human. Now if AI, by replacing some of these practical judgments, results in us practicing less we will, through our engagement with this technology, become less of ourselves. What is at issue is the judge who, like Forer, does less judging due to the introduction of sentencing guidelines, or the HR manager who does less hiring do to the algorithmic streamlining of her job. In each case the capacity for judgment is restricted and as a result the activities themselves become demoralized and thus demoralizing.[9]

[7] As Roochnik (2013, 170) puts it: “the *phronimos* accurately sizes up and then navigates effectively through the particulars of the situation.”

[8] This type of approximation raises some interesting questions internal to Aristotle’s function argument: if judgment is replaceable by machines we are not uniquely suited to it. If uniqueness is a criterion for identifying the human function, as Aristotle argues, the ability to rationally weigh particulars and make practical judgments about them may no longer be a viable candidate. Further if, for Aristotle, the acquisition of virtues is hierarchical and we move from habit based moral virtues to the more abstract intellectual ones, we may fail to develop the moral virtues in the first place because AI would diminish the need for judgment. There would, then, be a ripple effect on our capacity to display the intellectual virtues.

[9] The argument we offer here corresponds with the growing literature on deskilling – the impact of technology on the viability of important human capacities. Most relevant to our claims are Vallor (2016, section 3) and Danaher (2019b). Vallor produces a list of twelve techno-moral virtues (in reality, existing virtues that are reimagined to apply to the impact of technology). In part 3 of the book, she applies that list to contemporary technological questions ranging from the rise of social media to robotics and bio-enhancement. Vallor warns that these technologies threaten basic virtues and capabilities: too much engagement on our phones weakens social skills and our capacity to engage intimately. Autonomous weaponry may weaken a nation’s prudent hesitation before going to war. Caregiving robots may preclude meaningful opportunities for becoming empathetic or caring for those who depend on us. Extreme reliance on technology results in a loss

This is not necessarily an anxious, dystopian statement about technology's encroachment. Let us reiterate: There are good reasons to think that this kind of algorithmic decision making will yield good, ultimately better results, than our own judgment produces. It is quite possible that just as self-driving cars will ultimately become safer than human driven cars, algorithmic sentencing, police force allocation, and hiring decisions will ultimately be fairer and more efficient than what we have now. We are simply pointing out that the scope and opportunities for human judgment are narrowing. There is no reason to assume that synthetic judgments will be worse for being synthetic. Quite the opposite. But Aristotle tells us that we flourish by doing what human beings are naturally suited to do. And that is to use our judgment. AI encroaches on that territory. Now a critic may plausibly object to this argument along the following lines: Why is the concern about the atrophying of judgment such a big deal, why should we worry so much about our capacity for phronesis, especially if algorithmic replacements bring fairer, more equitable results for those who have been mistreated by our faulty, unfair, human, all too human judgments? Shouldn't we be concerned with better outcomes for those who were previously hurt by our misjudgments than about preserving the process of judging? Stated differently, perhaps the *virtuous* thing to do given the precarity and deficiency of our judgment is to gradually give it over to algorithms once it becomes clear they can do a better job than us? This is an important and powerful objection. But if we accept the Aristotelian premise that the ability to make judgments is a large part of what people value about themselves

of virtues centered on sociality, prudence, loyalty and commitment. Vallor's argument is related to the one offered here. But our focus is not so much on AI's impact on specific virtues as on its significance for the very conditions of becoming virtuous. If Aristotle is right that developing virtues turns on the capacity for phronesis, and if we are right that AI undermines phronesis, AI isn't just undermining or deskilling particular virtues – it is undermining the very possibility of developing virtues in the first place. Danaher argues that the rise of robots and AI will result in a shift from agency towards "patiency" – a passive, receptive mode of existing in which we are subject to and done to rather than responsibly originate action. Agency, Danaher argues, consists in four capacities: a "(i) the capacity for *sensing*, i.e. acquiring information from the world around us; (ii) the capacity for *processing*, i.e. categorising, sorting and rendering useful that information; (iii) the capacity for *acting*, i.e. the ability to use the processed information to form and implement action plans; and (iv) the capacity for *learning*, i.e. the ability to grow and develop the other three capacities." Robots and AI are meant to either supplement or replace all four. To the extent that they do, these technologies undermine agency. Aristotle's account of phronesis is not identical to what it means to be agentic but it is, of course, related. The phronimos must be able to acquire information, categorize and sort it, act on it, and learn how to improve in these three domains. Our Aristotelian focus is on the fourth aspect of agency. It is the learning, through the consistent, habitual exercise of practical judgment that ultimately gives rise to virtues. We are arguing that the opportunities for this kind of learning are being curtailed by the rise of AI.

(namely, they don't exclusively value better outcomes), then it is, at the very least, worth having a serious social discussion about whether the benefit of fairer outcomes is worth the cost of losing foundational capabilities. More descriptively, the paper's argument can serve as a reminder, for those who would embrace these improved outcomes, of the Rousseauian insight that technological progress rarely comes without a degree of regress. That the triumphalist picture of linear advancement suggested by the focus on better social results is over-simplified. That these improved results come with a significant diminution in what, until now, we assumed is an essential aspect of being a fully functional human. Second, if we accept the Aristotelian picture that tells us that the capacity for judgment is practice-dependent, that the ability to weigh particulars turns on having sufficient contexts in which to weigh them, the elimination of some of these contexts may diminish the capacity for judgment across the board. In other words, there may well be a spill-over effect from using one's judgment less in commercial and political contexts to the ability to use it in more private settings or in other areas that are not similarly influenced by machine learning.

4 Life Hacking

The risk for a decline in our capacity for practical judgment comes at the same time as artificial intelligence is making it possible to obtain and use more data about more aspects of our lives than ever before. Information about our sleep quality (based on a measurements of body temperature, room temperature, how many times we toss at night), heart rate, eating habits, exercise patterns, shopping behavior, heating and cooling preferences and so on, are now made available to us by a variety of apps deploying machine learning technology. An app called Sleep Cycle deploys a phone's sensors to figure out how much time we spend in each sleep cycle so that we can be woken at the optimal time. A sister app, Power Nap, is meant to facilitate an afternoon doze that keeps us out of deep sleep. Nest learns our home energy consumption patterns and adjusts our environmental systems to economize on costs.

Statistical machine learning based apps give us unprecedented amounts of data about our lives and come with the promise of life hacking – optimizing the way we sleep, eat, exercise, heat, cool and so on. There are, of course, remarkable benefits to be gained: It's good to save money on utilities, it may be helpful to know how our sleep cycles work, and it is critical to track one's sugar intake if one is diabetic.[10] But such tools also risk creating an excessive sense of

[10] The Seattle based company Brook provides such a service for diabetics, see https://www.brook.health.

self-importance – a technologically induced narcissism that convinces us that we are important enough to have all of this data collected about us. This is not the kind of self-importance that comes with the ability to flatteringly curate Facebook or LinkedIn profiles, or airbrush pictures on Instagram. There is an increasingly robust body of literature on how spending a great deal of time on social media carries the risk of exacerbating and sometimes even bringing about personality disorders including narcissism and OCD (see Aboujaoude 2012, Ch. 3; Rosen 2013; Terkle 2011). Machine learning driven life hacking is relatively new and its impact has not yet been rigorously studied. But it is plausible to assume that the massive collection of health information, behavioral information and other data points about the most humdrum aspects of our lives, with the purpose of hacking ourselves into optimal efficiency, will lead us to believe that we are the kind of creatures whose humdrum data matters a great deal. As AI helps us become preoccupied with how many minutes we spent in REM sleep, with how much protein we have ingested and with the variability of our heart rate through the day, it makes us into our own projects – rigorously and empirically engaging in tracking, assessing, and improving ourselves. A critic may object that, far from being problematic, such opportunities for self-design offer a calming antidote to the concerns about phronesis raised in the previous section, that life hacking provides us with a plethora of new contexts in which to exercise our judgment. And yet, all of these contexts pertain to ourselves. To develop practical judgment we must face outwards, towards others people and towards the world around us. The ability to navigate particulars well and to acquire virtues depends on exposure, on having a rich life in which one engages with others in a variety of situations in which our moral abilities are called on and challenged. An increased engagement with oneself does not qualify as such an existence.

The conjunction between a decline in the ability to use practical judgment on the one hand and an increased sense of self-importance on the other paints a worrying picture – we are at risk of taking ourselves more and more seriously even as our capacities for putting such evaluations in perspective decline. Self-love can be checked by sound practical judgment. But AI threatens a perfect storm: encouraging a growing preoccupation with ourselves while degrading the capacity to put such navel gazing into perspective.

5 Conclusion

If the right controls are put in place AI need not encode the prejudices of its creators. In fact, the technology is likely to become progressively fairer and safer.

The real reason to worry about AI is that it will gradually erode and displace our capacity for making practical judgments.

Techno optimists will scoff at this worry: technology always brings a degree of disruption and always generates the anxiety that the very essence of human nature is being undone. And yet here we still are. Automation, in particular, has always generated fears and anxieties about human displacement (see, e.g., Kang 2011; Mayor 2018). In fact, the very term luddite (used to describe those who are alarmed by the harms of new technology) commemorates early 19th century hosiery workers who smashed weaving machines for fear that the technology would put them out of work (see Randall 1998; Sale 1996). While the weaving machines probably did displace workers who could not acquire new skills, automation, more broadly, has tended to increase productivity rather than unemployment. This critique is sometimes referred to as the luddite fallacy (see, e.g., Ford 2009, Ch. 2).

But AI is inevitably bringing about a world with fewer contexts for making practical judgments. Some workers will benefit and move up the organizational hierarchy and will start making more sophisticated judgments than the ones that were just automated. AI will free a few of us up to be strategists. This is the usual, calming salve offered by the technology's proponents. When all the judgments that can be automated are automated, we will become available for more creative work. But this reply seems complacent; we can't all arrive at these sunlit uplands.[11] We are not all going to become strategists. Many important estimates suggest that within the next generation, a third or more of existing jobs will disappear due to AI powered automation.[12] In fact, the very attraction of the technology, the very reason that companies are engaged in an AI arms race, is the cost saving involved in the automation of the means of production. When AI works well we will be making the same number of products and we will be providing similar services with fewer people. This is its true economic draw (Roose 2019).

[11] A recent account of the impact of automation on Amazon fulfillment centers suggests that the jobs that were not automated became more rather than less dreary. Far from freeing remaining workers to perform more creative tasks, robots were introduced into the few parts of the work that afforded workers a degree of autonomy and privacy (such as walking through the warehouse to find ordered items). See MacGillis (2021).

[12] Job loss rates range between 14 and 47% percent in the next few decades. For a useful overview of key studies see: https://www.brookings.edu/blog/techtank/2018/04/18/will-robots-and-ai-take-your-job-the-economic-and-political-consequences-of-automation/. The mean rate is 38%. The World Economic Forum recently released a report that predicted that from the 1.37 million workers who will lose their jobs to automation in the next decade, only a quarter can benefit from programs that will teach them new skills. Three quarters will likely become unemployed. The report is available here: http://www3.weforum.org/docs/WEF_Towards_a_Reskilling_Revolution.pdf.

The techno-optimist will plausibly retort that less work does not necessarily mean fewer opportunities for developing or exercising judgment. Assuming we can financially sustain newly unemployed, and assuming we can create new sources of demand for automatically generated products in a market consisting of fewer paid workers, what is to stop people from honing their judgement in their leisurely activities? Why couldn't they develop phronesis by engaging in hobbies and artistic pursuits? Or while dedicating themselves more fully to family life?[13]

This is a potent objection. But like many kinds of philosophical optimism, techno-optimism suggests a problematic tradeoff in which we are to be assuaged about the concrete loss of existing goods by the promise of attractive yet hypothetical benefits. The future world of secure leisure may indeed provide remarkable opportunities for honing our judgment. But whether or not that world comes about depends on the prudent formulation and competent execution of public policy. In the meantime, the losses in skill and judgment are concrete and ongoing. What is the argument for preferring hypothetical future goods to foundational current skills? Further, there is reason to worry that, given leisure, future humans will not continue to hone their practical judgment and phronesis, practicing the foundations of their agency, but will passively slip into what (Danaher 2019a) and others (see, e.g., Floridi 1999; Gunkel 2011) describe as moral patiency. Stated differently, the techno-optimist argument that a world with less work will provide ample opportunities to develop phronesis depends on two separate layers of optimism: the first that circumstances and policy prowess will make widespread sustainable leisure practical, and the second that human psychology is robust enough to sustain agency and judgment under those new conditions.

The upshot of our argument is that while AI has the potential to streamline the functioning of both private and public enterprises and to reduce the biases inherent in their operations, the deployment of the technology to replace more and more everyday human judgments will undermine our capacity for judgment making. To the extent that this capacity is a foundational part of what we value about ourselves, this loss will adversely impact our well-being and diminish our self-understanding. Neither the utilitarians who claim that the tradeoff is worth it nor the techno-optimists who claim that it's not really a tradeoff offer convincing

[13] See, for example, Frude (2019). Danaher (2019b) provocatively proposes that a retreat into virtual worlds could help maintain key human capacities and allow us to flourish in a post-work world. Though automation can result in deskilling and a reduction of agency, these technologically created problems, Danaher argues, are susceptible to a technological solution.

arguments. A technology that threatens to undo a foundational human capacity deserves closer moral scrutiny.

References

Aboujaoude, E. 2012. *Virtually You: The Dangerous Powers of the E-Personality*. New York: Norton.

Annas, J. 1999 "Aristotle on Virtue and Happiness." In *Aristotle's Ethics*, edited by N. Sherman, 35–56. Lanham: Rowman & Littlefield Publishers.

Aristotle. 1999. *Nicomachean Ethics*. Translated by Martin Ostwald. London: Pearson.

Avati, A., K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shahet. 2018. "Improving Palliative Care with Deep Learning." *BMC Medical Informatics and Decision Making* 18 (Suppl. 4):122.

Awoyemi, J. O., A. O. Adetunmbi, and S. Oluwadare. 2017. "Credit Card Fraud Detection Using Machine Learning Techniques: A Comparative Analysis." In 2017 International Conference on Computing Networking and Informatics (ICCNI), https://doi.org/10.1109/ICCNI.2017.8123782.

Bostrom, N. 2016. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chawla, R. 2018. *How AI Supports Financial Institutions for Deciding Creditworthiness*. India: Entrepreneur. https://www.entrepreneur.com/article/310262.

Danaher, J. 2019a. "The Rise of the Robots and the Crisis of Moral Patiency." *AI & Society* 34 (1), https://doi.org/10.1007/s00146-017-0773-9.

Danaher, J. 2019b. *Automation and Utopia: Human Flourishing in a World without Work*. Cambridge: Harvard University Press.

Danks, D., and A. London. 2017. "Algorithmic Bias in Autonomous Systems." In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. IJCAI, https://doi.org/10.24963/ijcai.2017/654.

Domonoske, C. 2017. "Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk'." In *The Two-Way: Breaking News from NPR*, https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk.

Edwards, P. N. 1996. *The Closed World: Computers and the Politics of Discourse in Cold War America*. Cambridge: MIT Press.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Floridi, L. 1999. "Information Ethics: On the Philosophical Foundation of Computer Ethics." *Ethics and Information Technology* 1 (1): 37–56.

Ford, M. 2009. *The Lights in the Tunnel: Automation, Accelerating Technology, and the Economy of the Future*. United States: Acculant Publishing.

Fridman, L., D. E. Brown, M. Glazer, W. Angell, S. Dodd, B. Jenik, J. Terwilliger, A. Patsekin, J. Kindelsberger, L. Ding, S. Seaman, A. Mehler, A. Sipperley, A. Pettinato, B. Seppelt, L. Angell, B. Mehler and B. Reimer 2018. "MIT Autonomous Vehicle Technology Study: Large-Scale Deep Learning Based Analysis of Driver Behavior and Interaction with Automation." arXiv. https://arxiv.org/pdf/1711.06976.pdf.

Frude, N. 2019. "Technological Unemployment and Psychological Well-Being—Curse or Benefit?" In *In Education and Technological Unemployment*, edited by M. Peters, P. Jandrić, and A. Means. New York: Springer.

Gunkel, D. 2011. *The Machine Question*. Cambridge, MA: MIT Press.

Hughes, J. 2017. "Algorithms and Posthuman Governance." *Journal of Posthuman Studies* 1 (2): 166–84.

Kang, M. 2011. *Sublime Dreams of Living Machines: The Automaton in the European Imagination*. Cambridge: Harvard University Press.

Kearns, M., and A. Roth. 2019. *The Ethical Algorithm*. Oxford: Oxford University Press.

Kraut, R. 1989 *Aristotle on the Human Good*. Princeton: Princeton University Press.

Kuang, C. 2017. "Can AI be Taught to Explain Itself?" *New York Times Magazine*. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

Lohr, S. 2018. *"Facial Recognition is Accurate if You're A White Guy". New York Times*, https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html.

Love, D. 2014. *"Stephen Hawking is Worried about Artificial Intelligence Wiping Out Humanity" Business Insider*. https://www.businessinsider.com/stephen-hawking-on-artificial-intelligence-2014-5.

Mayor, A. 2018. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton: Princeton University Press.

MacGillis, A. 2021. *Fulfillment: Winning and Losing in One Click America*. New York: Farrar, Straus and Giroux.

McCarthy, J., M. L. Minsky, N. Rochester, and C. E. Shannon. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Scanned, annotated typescript, https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

Mindell, D. A. 1995. "Anti-Aircraft Fire Control and the Development of Integrated Systems at Sperry, 1925–1940." *IEEE Control Systems Magazine* 15 (2): 108–13.

Noble, S. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Pal, A. K., and S. Pal. 2013. "Evaluation of Teacher's Performance: A Data Mining Approach." *International Journal of Computer Science and Mobile Computing* 2 (12): 359–69.

Qin, C., D. Yao, Y. Shi, and Z. Song. 2018. "Computer-aided Detection in Chest Radiography Based on Artificial Intelligence: A Survey." *Biomedical Engineering Online* 17 (113), https://doi.org/10.1186/s12938-018-0544-y.

Randall, A. 1998. "The 'Lessons' of Luddism." *Endeavor* 22 (4): 152–5.

Randell, B. 1982. *The Origins of Digital Computers*. Springer.

Roche, T. 1988. "Ergon and Eudaimonia in Nichomachean Ethics: Reconsidering the Intellectualist Interpretation." *Journal of the History of Philosophy* (26): 173–94.

Roochnik, D. 2013. *Retrieving Aristotle in an Age of Crisis*. Albany: SUNY Press.

Roose, K. 2019. *The Hidden Automation Agenda of the Davos Elite*. *New York Times*. https://www.nytimes.com/2019/01/25/technology/automation-davos-world-economic-forum.html.

Rosen, L. 2013. *iDisorder: Understanding Our Obsession with Technology and Overcoming its Hold on Us*. New York: St Martin's Griffin.

Sale, K. 1996. *Rebels against the Future: The Luddites and Their War on the Industrial Revolution: Lessons for the Computer Age*. New York: Basic Books.

Schwartz, B., and K. Sharpe. 2011. *Practical Wisdom: The Right Way to Do the Right Thing*. New York: Riverhead Books.

Sumagaysay, L. 2018. “Less Biased Facial Recognition? Microsoft Touts Improvement, IBM Offering Help.” *The Mercury News*. https://phys.org/news/2018-06-biased-facial-recognition-microsoft-touts.html.

Terkle, S. 2011. *Alone Together Why We Expect More from Technology and Less from Each Other*. New York: Basic Books.

Vallor, S. 2016. *Technology and the Virtues: A Philosophical Guide to a World Worth Wanting*. Oxford: Oxford University Press.

Wang, P., R. Mathieu, and H. Cai. 2010. Predicting Criminal Recidivism with Support Vector Machines. In 2010 International Conference on Management and Service Science. https://www.researchgate.net/publication/251954270_Predicting_Criminal_Recidivism_with_Support_Vector_Machine.

Witten, I. H., E. Frank, and M. A. Hall. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington: Morgan Kaufman.