Humanistic Management Journal (2021) 6:355–371
https://doi.org/10.1007/s41463-021-00118-w

ORIGINAL RESEARCH

Master and Slave: the Dialectic of Human-Artificial Intelligence Engagement

Tae Wan Kim[1] · Fabrizio Maimone[2] · Katherina Pattit[3] · Alejo José Sison[4] · Benito Teehankee[5]

Received: 12 July 2021 / Accepted: 19 November 2021 / Published online: 3 December 2021

Abstract

The massive introduction of artificial intelligence (AI) has triggered significant societal concerns, ranging from "technological unemployment" and the dominance of algorithms in the work place and in everyday life, among others. While AI is made by humans and is, therefore, dependent on the latter for its purpose, the increasing capabilities of AI to carry out productive activities for humans can lead the latter to unwitting slavish existence. This has become evident, for example, in the area of social media use, where AI programmers tie psychology and persuasion to the human social need for approval and validation in ways that few users can resist. We argue that AI should serve humans with humans as masters and not the other way around. Moreover, we propose that virtue ethics might play a role to solidify the human as master of AI and guard against the alternative of AI as the master.

Keywords Human-artificial intelligence engagement · Virtue ethics · Human flourishing

Introduction

The massive introduction of artificial intelligence (AI) technologies has triggered significant societal concerns: AI-powered automation will replace humans and so-called "technological unemployment" will occur (Kim and Scheller-Wolf 2019); algorithms will direct workers at the workplace, functioning as algorithmic bosses (Lee 2016); consumers will increasingly defer to algorithms to make important decisions and eventually lose their autonomy and reflective agency to algorithms (André et al. 2018); due to sophisticated social robots that can satisfy social needs, people might prefer social robots to real humans, thus leading to loss of social and civic relationships and goods (Danaher and McArthur

✉ Benito Teehankee
benito.teehankee@dlsu.edu.ph

[1] Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA

[2] Department of Law, Economics, Politics and Modern Languages, LUMSA University, Rome, Italy

[3] Ethics & Business Law Department, University of St. Thomas, St. Paul, MN, USA

[4] Philosophy Department, University of Navarre, Pamplona, Spain

[5] Department of Management and Organization, De La Salle University, Manila, Philippines

2017); our growing reliance on algorithms for making important decisions will eventually turn democracy into “algocracy” (Danaher 2016). In short, in the AI dominated future, AI will be our master.

But perhaps visions of such a bleak future can be helpful to prompt reflection on alternative outcomes and what it might take to achieve them. Therefore, in this essay, we critically explore the role of AI in business and society in pursuit of a simple thesis: AI should serve humans with humans as masters and not the other way around. In particular we will focus on the role virtue ethics might play to solidify the human as master of AI and guard against the alternative of AI as the master. We will proceed as follows. Section 2 surveys and problematizes the massive use of AI in organizations and workplaces to delineate the challenge of the relationship between AI and humans. Section 3 introduces definitions of AI to help readers situate our main thesis, which will be developed in Section 4 with the dialectic of human-AI engagement and further elaborated in Section 5, which explains one mechanism of how humans might become subservient to AI. Section 6 discusses the role of virtue ethics in more detail, particularly its role in avoiding the AI-as-master “trap” and introduces ideas about how AI can be used to actually promote human flourishing.

AI, Work, and Organizations - the Potential for Optimized Decision Making

AI is one of the key technologies, or set of technologies, involved in the so-called digital transformation, i.e., the exploitation and progressive integration of digital technologies into the development of new products and services, production and business processes, supply chains and more generally into organizing and management processes (see Westerman et al. 2014; Matt et al. 2015). We can find intelligent computational systems in virtually all business functions: in the analysis of consumer needs and wants; the development of new products or services; interactions with customers; the organization of the entire production chain down to the support of individual decisions (Liao et al. 2017; Phillips-Wren 2012; Jarrahi 2018).

AI, then, is one of the key technologies of Industry 4.0 and enables the transformation of massive amounts of raw data into usable knowledge about decision and behavioral patterns (Sanders et al. 2016; Buer et al. 2018), particularly when it is integrated with other digital systems, such as the internet of things[1] and additive manufacturing (the process of creating an object by building it one layer at a time by using 3D printing) (Oztemel and Gursev 2020). While AI has contributed positively to cost reduction, quality improvement, sustainability (Kakhurel et al. 2018; Nishant et al. 2020), and improvement of working conditions and worker well-being by reducing repetitive tasks (see Bag et al. 2021), it is its role in decision support and knowledge management that has drawn attention. AI applied to individual, team and organizational learning may facilitate self-reflexive and reflexive learning and supports a wide range of functions that typically rely on the involvement and decision-making skills of humans. For example, AI can help in predictive maintenance of machine failures or it can automate customer support and relationship management by answering queries and analyzing opinions from clients’ correspondence. It can boost decision efficiency by managing e-mail or information retrieval from databases and can help

[1] https://www.oracle.com/it/internet-of-things/what-is-iot/

transform tacit to explicit knowledge through database mining for keywords, related concepts, and idea clusters. These remarkable positive contributions show how AI is seen as the driver of the "Fourth Industrial Revolution" (Schwab 2016). Before we take a critical look at the implications of the significant role of AI in human decision support - and even replacement - it is important to delineate what kind of AI we are focusing on in the context of our argument. Because the differences in AI systems are crucial to understanding the connection to virtue ethics, we ask the reader to bear with us as we delve into the nuances of AI and situate our argument.

Defining AI

Let us begin with a general definition: "artificial intelligence (AI) is a cross-disciplinary approach to understanding, modeling, and replicating intelligence and cognitive processes by invoking various computational, mathematical logical, mechanical, and even biological principles and devices" (Frankish and Ramsey 2014). Most important in this definition are the terms "replicating" and "intelligence," though definitions of these underlying concepts are inconsistent, referring to particular intelligent systems applied to specific domains (Hum AI Collab Key Insights; Carter 2018).

Pressed by the need to legislate, some governments such as that of the UK developed their own definition: "Technologies with the ability to perform tasks that would otherwise require human intelligence, such as visual perception, speech recognition, and language translation" (Carter 2018, 106). However, for at least two of these functions, it is not clear that human intelligence is strictly required, since even dogs are quite capable of visual perception and speech recognition. If intelligence means only human intelligence, the definition could be construed such that an AI technique is good to the extent that it replicates human intelligence. In particular, this definition implies that Deep Learning (DL) is not a good AI system because it does not replicate causal or counterfactual reasoning, a crucial aspect of human intelligence (Pearl and Mackenzie 2018). But if intelligence can mean animal intelligence, causal reasoning may not be necessary for a good AI system. There are successful robots, such as pet or military robots, that replicate non-human animals' cognitive capacity.

Therefore, perhaps an acceptable definition of AI is a combination of those offered by the Expert Groups of the Collaborations Between People and AI Systems (CPAIS) and the European Commission (AIHLEG): "any computational process or product that appears to demonstrate intelligence through non-biological/natural processes" (Annotation and Benchmarking…), "analyzing [its] environment --with some degree of autonomy-- to achieve specific goals" (2AIHLEG, 1). Not only is AI expected to perform intelligent functions, but also to change its environment within certain margins towards a preset orientation. Being non-biological or non-natural, the "artificial" in AI is clear. But the part of "intelligence" requires further examination.

In some sense "intelligence", first of all, denotes rationality, the abstract quality of doing things (or making things happen) with a view to an end or purpose, as opposed to chance. This entails an explanation, a propositional response to the question "why did this occur?" Today, AI typically refers to a certain kind of model that cannot be counter-factual and cannot logically reason with abstracts. So, it is in fact controversial whether AI is really intelligent. To deepen our understanding of the 'I' part of AI, we need to realize that there are two different competing models of AI: Machine Learning (ML)-based AI and Expert

Systems. The two kinds are distinct in the way that each understands and embodies “intelligence.” In cognitive science and philosophy, there are two competing and complementary theories of intelligence: Connectionism and Computationalism. Connectionism argues that the mind is a correlation engine, whereas computationalism argues that intelligence is a logical and symbolic reasoning system that understands causation and abstracts. ML-based AI is connectionist, whereas expert systems are computational and symbolic.

Connectionism argues that intelligence is primarily an end-to-end system, and its internal working is simply a bunch of correlations (Buckner and Garson 2019). This theory has been inspired by biochemical connections or neurons in the human brain, resulting in AI built on artificial neurons or neural nets. A major example of connectionist AI is Deep Learning (DL), a contemporary ML model which is a complex of non-linear, automated statistics, based on a myriad of hidden heuristics built on associations. The success of DL models relies on the availability of large amounts of data, more specifically training data, which is replicated and reinforced in the DL systems (Marcus 2018). This also means that if training data is biased or unethical, the outcome will replicate and reinforce biased or unethical patterns in outputs.

Computationalism argues that the human mind works like a computer, i.e. in accordance with transparent systematic abstract symbol-and-rule mechanisms that can be expressed with formal symbolic logic (Scheutz 2002). A typical use of computational AI is expert systems. Many successful expert systems consist of some kind of decision tree, which helps users to solve problems based on the background knowledge and logic input by developers. For instance, many autonomous vehicles use an expert system for its driving system because it uses codified rules from laws and local conventions to define safe driving. An advantage of symbolic or rule-based AI is the transparency of its internal working. Thus, when a developer finds a problem in the driving performance of an autonomous vehicle, she can see where the problems occurred and intuitively fix the problem. Of course, a disadvantage of an expert system is that it is not an automated learning system without human inputs, so it is time-consuming and labor-intensive to create. Another advantage of an expert system is that developers can teach it what a car ought to do by using abstract rules. In contrast, DL has difficulty learning rules. It learns from the exhibited behaviors of real human drivers.

Finally, to further clarify the nature of AI, let us add a commonly used distinction between weak and strong AI.[2] Strong AI is also called GAI (General Artificial Intelligence; or AGI). An example of general intelligence is a human who can solve problems across domains. Accordingly, GAI is also called “Human-like AI.” In contrast, weak AI is a domain-specific system. An AI system trained for translation is not able to drive a vehicle, but of course human translators can drive a car. There are groups of researchers specifically dedicated to studying and developing GAI who do not believe that realizing GAI is within our reach even in the coming century (Grace et al. 2018). Furthermore, the prediction concerns only domain-specific AI. Developing a single AI system that can automate all human jobs simultaneously is a totally different thing and so for our purposes we limit our discussion to weak AI. We also further focus more heavily on ML-based AI that is based on a connectivist view of intelligence because it currently dominates the space of technological advancements.

[2] Another way to divide strong and weak AI is that strong AI “seeks not only to think, but to feel and purpose as well, becoming a “mind” and not only a model of one, while “weak AI” is meant to be at the service of human designs (Botica 2017).

With the above definitions and delineations of AI we can now turn to the core of our essay, which is to examine the relationship between humans and AI. Above we have alluded to the different roles humans can have as trainers and designers of AI and we have introduced the various degrees of insight and understanding of the functioning of the resulting AI systems that humans can achieve. It is in these various combinations and relationships that the master and slave[3] dynamic between humans and AI is manifested.

Master and Slave: The Dialectic of Human-AI Engagement

The "Master-Slave Dialectic" in Hegel's Phenomenology of Spirit is perhaps the passage in his work that has been commented on the most. It has been used to explain a wide range of processes from how the human species evolved from lower life-forms ("hominization") and the psychological development of children, to the transformation of societies through industrialization and the history of nations as they progress into sovereign states. It could also serve as a lens through which to examine the trajectories of human-AI engagement.

The "Master-Slave Dialectic" is a conceptual construct, an idealized story of how two unequal individuals meet and experience a deep conflict or even life-threatening struggle in their joint quest for higher level self-consciousness. As they go through different stages in their relationship, they come to realize how they inescapably depend on each other. Superior self-awareness could only come through recognition of and from the other; self-reflection could only be achieved through the mediation of the other as a mirror. Although to affirm the self, one needs to deny the other, at the same time the other turns out to be necessary, even if just to forge the notion of self. Despite the inequality, there is mutual and reciprocal need between them.

AI represents the latest episode in the grand scheme of technological advancements. Like all tools, whether as automated algorithms alone or as algorithms embedded in robots, AI was invented by humans to make work easier, thus allowing us to save time and make life more pleasant by freeing it from drudgery. AI accomplishes this not only by assisting in routine tasks, but also by augmenting and enhancing human agency (for instance through precise medical imaging diagnostics). In that respect humans are masters, and AI is our close-to-ideal servant or serf. Indeed, the term "robot" comes from the Czech word for "serf", a tenant who pays rent on farmland through servitude or labor. Yet there is a constant danger of dependence, exacerbated by the fact that AI can potentially do tasks such as driving, previously imagined to be exclusive to humans. It is at this juncture that Hegel's dialectic becomes highly relevant. By learning to perform activities in ways even better than humans can, might there be a chance that AI one day will be our master and we will be the serf? While we might not know the answer to that question, we do know that AI ought to not be the master.

There are certainly limits to Hegel's master-slave allegory applied to human-AI relationships. First, despite the original inequality, the master did not create the slave, rather, they just found each other almost by chance, as previously existing individuals. Humans, on the other hand, created AI from scratch, although they are unable to endow it with life. AI, currently, fully depends on humans for existence, its goals set and determined by humans.

[3] It should be noted that we deliberately use the term "slave" in this essay in reference to Hegel's use of the term in the "Master-Slave Dialectic", without connecting to the practice of slavery, past and present, or reference to persons with dignity that were or are enslaved.

Between humans and AI there is no intersubjectivity. No matter how expert and efficient AI can become, surpassing humans in particular tasks, it can never establish its own purpose; it has no preferences or desires; it experiences no satisfaction or fulfillment. The ends of AI will always be extrinsic; it is a serf by "nature".

The slavish nature of AI, however, can turn into mastery via "nurture." In the Hegelian version, inexorably, the master becomes dependent on the slave, even as the slave, in turn, becomes dependent, not only on the master, but also on nature as the source and store of raw materials for its work. By carrying out productive activities for the master, however, the slave develops intelligence, skills and creativity, grounds for recognition and heightened self-knowledge. Meanwhile, the master, by contrast, regresses to a life dedicated to consumption and enjoyment, becoming no different from individuals belonging to irrational life-forms. The lack of work has pushed the master unwittingly to a slavish existence.

Troubling as this twist of fate may seem, it is far from inevitable. Not being alive, AI will not develop intelligent consciousness. Therefore, the actual danger lies not with AI, but with humans themselves as they regress into a state of consumption and enjoyment. There are instances in which dependence on AI is not a bad thing at all: think of robot bomb-defusers. Yet admittedly, there are occasions in which over-dependence on AI causes distinctive human powers to atrophy. Why bother to remember telephone numbers, birthdays, addresses, and the like when there are voice-controlled PDAs like Siri, Alexa, or Cortana? What's the point in learning math, memorizing poetry, or even learning a language with Google's ever-expanding array of apps?

In the space between benefits from AI and harm to human flourishing, the role of virtues, in particular temperance, is starting to become clear. Humans ought to make every effort to retain dominion and find moderation between trying to be spared immediate pain and suffering, but not at the steep price of losing higher-order agency and fulfillment. Since AI will never refuse to do the human's bidding, AI will only be as good as its human master and commander. So far, we have drawn a more general comparison between the Hegelian master-slave relationship and the potential devolution of the human into the serf of the AI. In order to flesh out how this might manifest we will next take a closer look at the way AI affects humans in the example of social media, which is an area AI has been used extensively in support of the business model.

How AI-Enabled Social Media Can Harm Well-Being

Given that mastery of moderation or temperance as a virtue requires practice and reflection, it is important to lay out systematically in which way human flourishing can be affected in the context of ubiquitous AI-enabled technologies, such as smartphones and social media (Brendel et al. 2021; O'Neil 2016). Because these technologies are often thought of as neutral tools, they may escape critical evaluation in terms of the master-slave dialectic. It is telling that Facebook's CEO acknowledged in a 2018 social media post that neutrality of this particular AI application cannot be maintained and that a balance between benefit and harm is important:

> We feel a responsibility to make sure our services aren't just fun to use, but also good for people's well-being. … The research shows that when we use social media to connect with people we care about, it can be good for our well-being. We can feel more connected and less lonely, and that correlates with long term measures of happiness and health. On the other hand, passively reading articles or watching videos --

> even if they're entertaining or informative -- may not be as good. (Mark Zuckerberg, January 12, 2018, https://web.facebook.com/zuck/posts/one-of-our-big-focus-areas-for-2018-is-making-sure-the-time-we-all-spend-on-face/10104413015393571?_rdc=1&_rdr)

This was the first time Facebook had acknowledged the possibility that prolonged use of the platform may be harmful to humans. Zuckerberg mentions the positive potential of social media as a tool for building social connections and thereby improving happiness and health. However, emerging problems with Internet addiction and self-esteem among young people had been reported by researchers in relation to popular interactive Internet sites around the same time as the birth of Facebook in the mid-2000s (Widyanto and Griffiths 2006) In this regard, Zuckerberg is less forthcoming. It is not how users "passively" consume articles and videos on the platform that is the problem. In fact, the Facebook business model rests on users not being passive at all but in their active interaction with media through commenting, liking, sharing, clicking, and tagging (Tabaka 2017). But to understand the insidious impacts of prolonged social media use, one needs to look at how social media sites use AI to encourage ever-increasing use and active engagement among people (McNamee 2019; Ryan et al. 2014; Vallor 2016).

The main mechanism by which this is achieved directly derives from behavior modification techniques developed by psychologists and surreptitiously deployed through big-data enabled algorithms—i.e., AI/ML. Thus, social media AI programmers combine psychology and persuasion concepts from the early twentieth century, like propaganda, with techniques from slot machines (e.g. variable rewards), and tie them to the human social need for approval and validation in ways that few users can resist. Stanford professor Fogg (1996) coined the term "captology" to describe this mechanism which has been developed in perhaps its highest form by Facebook. The key implication for human flourishing here is the loss of autonomy because of behavioral manipulation. Tristan Harris (2015, 2017), former Google design ethicist and president and co-founder of the Center for Humane Technology referred to it as "brain hacking"— one of the major mechanics through which humans slip into the role of slaves (Tabaka 2017; Bosker 2016).

Aside from its addictive impacts, social media has also been associated with increased stress levels. Morin-Major et al. (2016) found evidence that Facebook behaviors are associated with cortisol concentrations in adolescents during the day. Cortisol is the body's leading stress hormone and high levels cause feelings of anxiety among people. Such feelings are compounded by social envy and depression, which are also associated with extensive social media use. Appel et al. (2016) reviewed several studies on this topic and found that passive Facebook use indeed predicts different measures of social comparison as well as envy. In several studies, social comparison or envy mediate a positive association between Facebook use and undesirable affective outcomes such as depression. While the causality still needs further study, it is ironic that we associate social media with the positive impacts of social connections but because of how the platform is used, which is mainly to project positive life events, the unintended consequence of envy and depression may be triggered. The user's "liking" of others' posts is used by algorithms to show more of the same types of posts to the user, triggering the belief that others' lives are better than one's own. Furthermore, the tendency of social media to trigger misunderstandings and polarized filter bubbles also limits effective discussions of complex issues (Pariser 2012).

While we have used Facebook as a concrete example, and a well-studied one, this particular technology is not the only cause of harm. Almost all companies that use AI technologies rely on similar tactics to influence consumers and workers. And the kinds of harm caused by them facilitate the development of "slavedom." In light of this expanding threat to our "mastery" and flourishing more broadly, we propose that we approach the interaction with and design of AI through virtue ethics, which we discuss further in the next sections.

How Virtue Ethics Can Help us to Be Masters

Approaches to Evaluating the Ethics of AI

AI has been described above as processes or products that imitate human intelligence and thus it is challenging to examine the ethics of AI because ethics is concerned with what's right and wrong in human action. AI isn't human, although it's human-made, and it only imitates, and does not actually perform, intelligent human activity. Therefore, like all tools or machines, AI can only be appraised technically, whether it produces the desired output (effectiveness) and whether it does so in a manner that optimizes the use of resources (efficiency or economy), but not ethically. It is not enough to be a "functional equivalent" of human action to be subject to ethical judgment; agency itself has to be human, that is, proceeding freely and purposefully from an individual belonging to the human species.

There can only be room for ethical judgment, for moral praise or blame, in the way humans engage with AI. Just like all artifacts, humans make use of AI in order to augment or enhance their own activities, such that AI somehow extends, but never entirely supplants human agency. Through the use of AI, we could program a machine to emit sounds similar to human speech, but that can only happen thanks to our inputs, even when the resulting outputs are to some extent unforeseen or novel. AI cannot come up with completely original speech inasmuch as it depends on previous data and algorithms which identify statistical correlations among them. That is why ethical judgment always bears upon human agents, never on AI itself.

More specifically, humans develop, deploy and use AI, and oftentimes, with a business intent or purpose. This is the precise subject matter of our ethical investigations. While acknowledging its socially transformative and revolutionary potential, we shouldn't forget that "AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good" (HLGAI 2019, 4). Due to the central role of human flourishing in the ethical judgement about AI, we need to define the scope of human flourishing. As we will further discuss below, the role of virtue and ethical development is cardinal in flourishing. But we do not deny that we need other dimensions of integral human development such as health, intellectual, social, aesthetic, emotional and spiritual (Alford and Naughton 2001). Once we clarify all the dimensions, we can map how AI can contribute to each of these dimensions.

Engagement with AI in business is to be deemed ethical insofar as it contributes to the common good of flourishing, both of the individual and of society as a whole. The decision to use AI in business is usually taken in accordance with teleological or utilitarian principles, broadly construed, after an analysis of the costs and benefits of the alternative. However, most approaches to AI ethics follow the deontological school, concerned above all with safeguarding fundamental human rights. A prime example is the document "Ethical Guidelines for Trustworthy AI" sponsored by the European Commission (AI HLEG 2019). The Guidelines mandate that AI respect human autonomy, prevent harm, uphold fairness, and remain explicable. In particular, throughout its life-cycle, AI should pay attention to seven key requirements which spell out the above: support human agency and defer to human oversight; be technically robust, preventing or minimizing harms to human integrity; protect data privacy; be transparent in data management and decision making; allow for diversity and inclusion, eschewing unfair discrimination; preserve societal and environmental wellbeing; and exhibit accountability.

Although at first there seems to be no difficulty with the rights and values that deontological approaches towards AI present, in fact they offer little practical guidance to navigate the conflicts and tradeoffs that invariably surface in human-AI engagement. For instance, privacy and security demands that sensitive information provided by users or generated by them in interaction with AI (preferences, sex, age, religious or political views, and so forth) not be used unlawfully and be accessible only by authorized agents. Yet at the same time, AI transparency and explicability requires precisely that data gathering, labelling, and processing be documented, so as to allow traceability of possible errors and biases in decisions.

Mainly due to this shortcoming in dealing with tensions and trade-offs between rights and values, we turn to explore the somewhat less common virtue ethics approach. Our purpose isn't so much to replace, but to extend the deontological method, together with its rules on human dignity, harm prevention, and fairness. However, before explaining how virtue ethics can go beyond deontology in dealing with conflicting principles of human-AI engagement, we shall first explain other fundamental aspects in which the two differ.

Virtue Ethics and AI

Virtue ethics, in contrast with the deontological approach, focuses on the agent, not the action and its conformity with rules. It considers how agents achieve the moral good proper to their nature, namely flourishing, by cultivating the distinctive human excellences or virtues through free and purposeful actions. Human engagement with AI is just one among the myriad of opportunities to develop the virtues which are partially constitutive of their final end of flourishing.

The question now is how to develop the virtues while engaging with AI. There have been a couple of attempts to explain this centering on the intellectual virtues (including practical wisdom, which is in part intellectual, in part moral). Since AI deals mainly with data, information, and statistical correlations for decision making, this emphasis is highly understandable. For Grodzinsky (2017: 222), Big Data is essentially a quantitative measure of human behavior to which AI is applied to perform predictive analyses on the basis of correlations. AI has shown exceptional ability in interpolation, that is, predicting what happens next if a trend continues; not so in the case of extrapolation, when no such trend has been identified. In principle, correlations discovered by AI can point out possible causation, but only humans can take that leap, taking advantage of AI leads and making use of other complementary scientific methods.

Ironically, AI was invented in part to remedy human weaknesses, both intellectual and moral, in deliberation, decision making, and action. For instance, in selecting candidates for employment, we would like to be free from biases all humans have to some degree regarding sex, age, race, and so forth, zeroing in on the best individual on the basis of predetermined desirable characteristics. But AI systems need and depend on (historical) data and algorithms, provided and generated by biased human beings. As a result, AI becomes quite useless in eliminating biases and instead serves to extend and perhaps even augment them.

One approach is to ignore ethics altogether and simply to use all data, accept their messiness, and focus exclusively on finding correlations among them (Mayer-Schonberger and Cukier 2013). But one could argue this position is itself unethical, granted that humans are inescapably ethical beings, who reflect their values and moral worth through freely chosen actions. Another is to acknowledge the ethical import of human-AI engagement.

Grodzinsky (2017) chooses the latter and strives to discover the intellectual virtues or excellences Big Data scientists need to perform their work well. She states the discovery of statistical correlations cannot be the sole epistemic end of Big Data research. Rather, there should be room likewise for other types of knowledge and greater understanding of phenomena (for instance, through causal relations), not only for possible practical applications, but also in themselves. This implies recognizing that data is never free from hypotheses or background theories and human biases. Data are responses to questions containing complex beliefs or intuitions that are incomplete or require verification. Drawing attention to certain data instead of others denotes external interests and ulterior motivations on the part of researchers. Moreover, data always need a theoretical framework to make sense, be interpreted, or understood as they are never isolated or self-contained. Communicating data meaningfully requires a narrative thread. And in order to draw useful inferences from data, we need not only domain-specific knowledge, but also a worldview and a commitment to values only human beings can supply. Echoing Douglas (2009), Grodzinsky (2017: 228) affirms that values may not serve the same role as evidence in scientific research, but they complement it.

What are the intellectual virtues, i.e., the acquired habits of thought, data scientists need to perform their job well? Grodzinsky offers a long list: creativity, curiosity, critical thinking, collaboration, communication, humility, prudence, intellectual courage (Grodzinsky 2017: 229). However, in the end, she hones in on three (Grodzinsky 2017: 233). First, open-mindedness in taking generated patterns and predictions and putting them into context, although they might seem counterintuitive to initial hypotheses. Second, rigor in validating the evidence of predications, producing reliability and trust. And third, honesty in documenting and communicating findings, so as to ensure transparency, distribute responsibility adequately among all agents, and safeguard the openness of data. Only thus will data scientists behave responsibly in their practice as members of a community, reliably evaluating inputs to models and models themselves (algorithms, variables, data sets), together with the correlations and patterns that emerge from their study.

However, intellectual virtues are not the only virtues as there are moral ones that are equally relevant for proper human-AI interaction. Concern over the moral virtue of practical wisdom has arisen over the challenges that automation (through machine learning and robotics) poses to human work (Vallor and Bekey 2017). For not only can human work be facilitated, augmented, or enhanced, but it can also be substituted or replaced by AI systems. As a result, there is a loss of work for humans (even fear that there may not be enough work AI *cannot* do) and a loss of wages, as humans fail to compete with AI productivity. In an extreme version, we are before the "end of work" dystopia.

The need for responsible self-regulation in light of holistic long-term values as practical wisdom affords, nonetheless, will not disappear even in a world where AI is omnipresent. In recent times, studies have underscored the emotional roots of moral experience, something to which AI is absolutely impervious. In fact, part of AI's advantage over humans was precisely this, its lack of feelings. AI was meant to overcome human weaknesses without actually remedying, but rather by sidestepping them. For example, machine translation saves humans the trouble of learning a language (the weakness) while allowing them to communicate (the solution). Unlike AI, the virtues actually remedy human weaknesses or failings at their root by helping to learn a language, for instance, to follow through with the example above.

Practical wisdom is not mere technical expertise that uses the most adequate means to a given end. It entails the choice of the right end, besides, in light of which one decides on the means. Further, it includes an all-encompassing or moral evaluation of the acting self with regard to the end and means chosen. Vallor and Bekey (2017) unpacks the distinctiveness of practical wisdom in comparison to AI "substitutes" in three components. First, practical wisdom allows for decision making on complex goals, over the span of a lifetime (long term), all things considered. Second, practical wisdom permits one to identify an ultimate goal or final end among several incompatible options, while providing reasons not only to oneself but also to other affected parties, that is, intersubjectively. Third, practical wisdom involves taking ownership or responsibility over decisions and self-regulation toward a freely chosen end goal, with which definitive "success", the good, or absolute perfection is measured.

Practical wisdom cannot be reduced to the productive expertise AI promises or delivers. No matter how sophisticated or capable, AI systems are mere extensions of ourselves, depending on us for their own existence and maintenance, even though we may not fully understand their decision-making processes. AI systems cannot take responsibility for themselves or their interventions, something which remains the sole prerogative of human beings. However, practical wisdom is not the only moral virtue humans can exercise in engagement with AI. Justice, courage, and moderation, the so-called "cardinal virtues" insofar as they act as hinges on which all the other moral virtues rest, also enter into play.

AI scientists and users need justice in order not to exacerbate historic inequalities in employee selection and criminal sentencing software, for instance. Justice also requires special attention to the most vulnerable, such as the poor, children, senior citizens, the disabled, and the marginalized, so that they likewise have access to AI and participate in its benefits. For example, the visual or hearing impaired could take advantage of AI systems in order to navigate through cities or have access to data or information through specially designed interfaces.

Courage is equally necessary so as to achieve optimal AI use. Instead of putting brakes on innovation, AI research should be encouraged and promoted despite myriad difficulties. It would be a great blessing for humankind if surface travel through self-driving vehicles, combinations of AI and robotics, where to attain the safety standards of air travel, to cite an example. Similarly, the extensive use of AI in medical diagnostics could boost early detection of illnesses and improve prognosis. And AI embedded in hardware could take over hazardous jobs in mining or bomb-defusing, for instance.

Moderation too is required in order to ensure AI fulfils its potential in contributing to flourishing. On the one hand, marketers of products that create addictions or dependencies should take the proper measures in touting them to vulnerable populations. On the other hand, governments must also take care in their data gathering efforts so as not to intrude into the privacy of citizens without warrant, becoming in effect surveillance states. Otherwise, both marketers and governments would fail to respect the dignity of customers and citizens.

Figure 1 illustrates the role the five virtues introduced above play in maintaining the proper ordering between AI and humans: keeping humans firmly in the position of mastery with AI being the servant. Beyond the positive impact on flourishing that can be secured through this proper ordering, AI can further assist humans in the process of

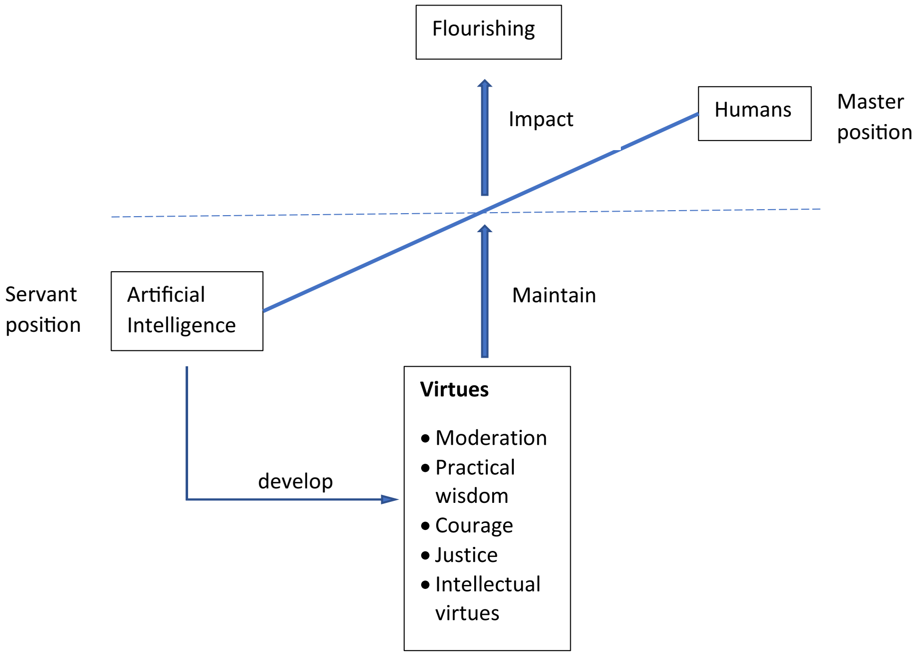

Fig. 1 The role of virtue ethics in the AI-human relationship

self-improvement and become perfective by allowing reflection and insight into human weaknesses, which we discuss below in the context of practical examples.

AI as a Tool to Improve ourselves

One area of human "imperfection" that appears often in the context of AI use - and abuse - are biases. While one of the most common ethical challenges of using AI is the perpetuation of such human biases (Angwin et al. 2016; McInnis et al. 2016; Olteanu et al. 2019; Raji and Buolamwini 2019), use of artificial intelligence has actually helped lay bare and provide concrete evidence of how such biases manifest. This in turn has provided opportunities for individual growth as well as organizational growth (Florentine 2016; Kleinberg et al. 2018). By showing the outcomes of a large number of decisions made by AI that are based on demonstrated individual decisions, it can become visible and concrete how the totality of incremental bias results in harmful treatment. Furthermore, when explicit processes are put into place to examine datasets before AI uses them to develop a decision function, data mining facilitated by AI can also help to mitigate bias if it is identified as being present in the dataset (Vasconcelos et al. 2018). If using AI can help us see who we are and what we are becoming, it offers a chance to examine whether this is who we want to be.

MIT's moral machine (https://www.moralmachine.net/) is one example of an AI-related "game" that allows individuals to examine their own biases and/or specific value systems that might be unconsciously at work in their decision making. This interface asks people to decide about a variety of moral dilemmas a machine might be facing

– for example a self-driving car with failing brakes and humans in harm's way inside and outside of the car. In a series of scenarios, one is prompted to choose whether the car should proceed straight ahead harming pedestrians in a crosswalk (which in various scenarios can be old people, men, women, children, robbers, etc.) thus saving the passengers (also a variety of people depending on scenario) or the other way around. By presenting a number of different combinations, such as harming 5 old people versus saving one child, and forcing choice, the moral machine thus reveals the implicit values that guide the decision maker. Maybe one finds out that one places more value on men than women or young over old persons. MIT's moral machine has also aggregated the results of everyone participating in this exercise thus allowing comparisons across groups of persons. Individually we can thus learn something about ourselves, but we can also learn about commonalities and differences in values and biases we might have with others.

Another example of bias being laid bare by engagement with algorithms is "survival of the best fit" (www.survivalofthebestfit.com), which is an educational game about hiring bias based on machine learning. It demonstrates how AI can inherit human biases and further inequality. In this game one takes on the role of the owner of a startup venture needing to make hiring decisions based on seeing "CV"s with various combinations of skills, school prestige, work experience and ambition of each applicant. Each round of hiring increases the time pressures resulting in the introduction of algorithms based on one's own hiring decisions done manually as well as decisions about which data sets to use in creating an algorithm. In various steps the game shows in which way one's own decisions were biased and how the introduction of machine learning resulted in biased hiring, despite good intentions by the player.

These two examples are perhaps games more than actual AI but they have been developed to help people reflect on the way AI can and does amplify the very imperfect human biases that are inevitably present in every decision we make (Banaji and Greenwald 2016). This is particularly the case when such biases are extrapolated and magnified, which is exactly the purpose of these games. While AI use can exacerbate social injustices, the introduction and use of AI – with detrimental results – is opening the door to self-examination and reflection, which in turn allows for improvement. The display of the harmful results of multiplication of individual bias thus is a magnifying glass through which we can better see ourselves the way we truly are.

Another way AI can highlight areas of human behavior that might need reflection and work to improve is its function as an objective and impartial mirror that can provide feedback about how we behave which we might not be able to see (or which we do not want to see). In that sense it can play the role akin to a therapist that points out as a trained observer what might be going on in our lives and causing suffering and pain. Take for example the experiment of Microsoft's chatbot "Tay", which was supposed to be responsive to people's messages and be able to carry on a casual and playful conversation. "Tay, instead of enhancing 'her' linguistic fluency by navigating the Internet space, turned into a representative of the more horrific face of social media and adopted a chaotic, crudely sexist and racist (anti-Semitic) mode of talk" (Beran 2018). Within less than 24 h the bot developed into a racist, misogynistic and generally horrid conversation generator (Bird et al. 2018; Wakefield 2016). By being programmed to mirror behavior it encountered, the bot showed very clearly what the tenor of conversation in social media – here Twitter – actually looks like. Of course, it is not like people were not aware that short online speech bursts do not bring about the best in persons, but having an impersonal and objective AI enabled bot whose very design was focused around mirroring speech, is a powerful mechanism to prompt reflection.

Conclusion

AI systems are instruments or tools invented for the ultimate purpose of contributing to flourishing, the good life for human beings in society. Rules or behavioral norms are necessary to ensure proper human-AI engagement, particularly in developing or designing, deploying, and using such systems. AI can augment, extend, and enhance human agency in perception, reasoning or decision making, and actuation.

In this essay we have laid out a thesis that whether AI can continue providing these benefits in light of significant potential harms to individual and societal well-being is dependent on maintaining the proper relationship between AI and humans. Drawing on Hegel's Master-Slave dialectic, we have highlighted the dangers to flourishing when humans cannot maintain mastery over AI but rather develop into becoming AI's servants. In order to better understand the dangers from an inverse AI-human relationship and also find approaches to avoid them, we have proposed a virtue ethics lens. A virtue ethics approach can provide concrete suggestions for the issues that should be considered in the design and use of AI systems where other perspectives, such as teleological or deontological approaches, might fall short.

In particular, as Fig. 1 illustrates, we draw attention to five virtues that should be at the center of reflection about how humans can retain their position of mastery over AI. These 5 virtues, if kept in the forefront when designing, using, and advancing AI technology, can help maintain the proper ordering between AI and humans - indicated in Fig. 1 with an upward sloping line between them. In such a properly ordered relationship between AI and humans, AI can actually have a positive impact on the development of virtues as well.

The virtues lens for approaching the relationship between AI and humans that has been developed in this essay is focused on weak, machine learning-based forms of AI. We have set this narrower lens because the use of this type of AI is widespread and because it's contributions and dangers to flourishing are well documented. However, this lens can also be applied to evaluating the development and use of strong AI systems. In fact, we believe that bringing a virtues perspective to understanding the relationship between humans and strong AI is even more important because of the increasing removal of humans from the decision making and action of strong AI systems. The discussion of the role of virtues in strong AI systems is therefore an important area for future study.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest Not applicable.

References

Alford, Teresa, and Michael Naughton. 2001. *Managing as if faith mattered*. Notre Dame: Notre Dame University Press.

André, Quentin, Ziv Carmon, Klaus Wertenbroch, Alia Crum, Douglas Frank, William Goldstein, et al. 2018. Consumer choice and autonomy in the age of artificial intelligence and big data. *Customer Needs and Solutions* 5: 28–37.

Angwin, Julia, Jeff Larson, Surya Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. ProPublica. Retrieved December 3, 2021 at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Appel, H., Alexander Gerlach, and Jan Crusius. 2016. The interplay between Facebook use, social comparison, envy, and depression. *Current Opinion in Psychology* 9: 44–49. https://doi.org/10.1016/j.copsyc.2015.10.006.

Bag, S., J.H.C. Pretorius, S. Gupta, and Y.K. Dwivedi. 2021. Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities. *Technological Forecasting and Social Change* 163: 120420.

Banaji, Mazarin R., and Anthony Greenwald. 2016. *Blindspot: Hidden biases of good people*. New York: Bantam.

Beran, Ondřej. 2018. An attitude towards an artificial soul? Responses to the "Nazi Chatbot". *Philosophical Investigations* 41: 42–69.

Bird, Jordan J., Anikó Ekárt, and Diego R. Faria. 2018. *Learning from interaction: An intelligent networked-based human-bot and bot-bot chatbot system. UK workshop on computational intelligence*. Cham: Springer.

Bosker, Bianca. 2016. The binge breaker. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2016/11/the-binge-breaker/501122/

Brendel, A.B., M. Mirbabaie, T.B. Lembcke, and L. Hofeditz. 2021. Ethical management of artificial intelligence. *Sustainability* 13: 1974.

Buckner, Cameron and James Garson. 2019. Connectionism. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 edition). Metaphysics Research Lab, Stanford University.

Buer, Sven-Vegard, Jan Ola Strandhagen, and Felix Chan. 2018. The link between industry 4.0 and lean manufacturing: Mapping current research and establishing a research agenda. *International Journal of Production Research* 56: 2924–2940.

Carter, D. 2018. How real is the impact of artificial intelligence? The business information survey 2018. *Business Information Review* 35 (3): 99–115.

Danaher, John. 2016. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 29: 245–268.

Danaher, John, and N. McArthur, eds. 2017. *Robot sex: Social and ethical implications*. Cambridge, MA: MIT Press.

Douglas, Heather. 2009. *Science, policy, and the value-free ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Florentine, Sharon. 2016. How artificial intelligence can eliminate bias in hiring. https://www.cio.com/article/3152798/artificialintelligence/how-artificial-intelligence-can-eliminate-biasin-hiring.html.

Fogg, B.J. 1996. *Persuasive technology: Using computers to change what we think and do*. San Francisco: Morgan Kaufmann.

Frankish, Keith, and William Ramsey. 2014. *The Cambridge handbook of artificial intelligence*. Cambridge: Cambridge University Press.

Grace, Katja, et al. 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research* 62: 729–754.

Grodzinsky, F.S. 2017. Why big data needs the virtues. In *Philosophy and computing essays in epistemology, philosophy of mind, logic, and ethics*, ed. T.M. Powers, 221–234. Berlin: Springer.

Harris, T. (2015). How Technology is Hijacking Your Mind — from a Magician and Google Design Ethicist. *Thrive Global*. https://medium.com/thrive-global/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3

Harris, T. (2017). Our minds have been hijacked by our phones. Tristan Harris wants to rescue them. *Wired*. https://www.wired.com/story/our-minds-have-been-hijacked-by-our-phones-tristan-harris-wants-to-rescue-them/

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics guidelines for trustworthy AI*. Brussels: European Commission.

Jarrahi, M.H. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons* 61: 577–586.

Khakurel, J., B. Penzenstadler, J. Porras, A. Knutas, and W. Zhang. 2018. The rise of artificial intelligence under the lens of sustainability. *Technologies* 6 (4): 100.

Kim, T.W., and A. Scheller-Wolf. 2019. Technological unemployment, meaning in life, purpose of business, and the future of stakeholders. *Journal of Business Ethics* 160: 319–337.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass Sunstein. 2018. Discrimination in the age of algorithms. *Journal of Legal Analysis* 10: 113–174.

Lee, Min Kyung. 2016. Algorithmic bosses, robotic colleagues: Toward human-centered algorithmic workplaces. *XRDS: Crossroads. The ACM Magazine for Students* 23: 42–47.

Liao, Yongxin, Fernando Deschamps, Eduardo de Freitas, Rocha Loures, and Felipe Pierin Ramos. 2017. Past, present and future of industry 4.0-a systematic literature review and research agenda proposal. *International Journal of Production Research* 55: 3609–3629.

Marcus, Gary. 2018. Deep learning: A critical appraisal. arXiv preprint 1801.00631.

Matt, Christian, Thomas Hess, and Alexander Benlian. 2015. Digital transformation strategies. *Business & Information Systems Engineering* 57: 339–343.

Mayer-Schonberger, Viktor and Cukier, Kenneth. 2013. Big Data: A Revolution That Will Transform How We Live, Work, and Think

McInnis, Brian, Dan Cosley, Chaebong Nam and Gilly Leshed. 2016. Taking a hit: Designing around rejection, mistrust, risk, and workers' experiences in Amazon mechanical Turk. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems (CHI 2016*). ACM.

McNamee, Roger. 2019. *Zucked: Waking up to the Facebook catastrophe*. New York: Penguin.

Morin-Major, Julie Katia, Marie-France Marin, Nadia Durand, Nathalie Wan, Robert-Paul Juster, and Sonia Lupien. 2016. Facebook behaviors associated with diurnal cortisol in adolescents: Is befriending stressful? *Psychoneuroendocrinology* 63: 238–246. https://doi.org/10.1016/j.psyneuen.2015.10.005.

Nishant, Rohit, Mike Kennedy, and Jacqueline Corbett. 2020. Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management* 53: 102104.

O'Neil, Cathy. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

Olteanu, A., C. Castillo, F. Diaz, and E. Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2: 13.

Oztemel, Ercan, and Samet Gursev. 2020. Literature review of industry 4.0 and related technologies. *Journal of Intelligent Manufacturing* 31: 127–182.

Pariser, Eli. 2012. *The filter bubble: How the new personalized web is changing what we read and how we think*. New York: Penguin.

Pearl, Judea, and Dana Mackenzie. 2018. *The book of why: The new science of cause and effect*. Basic Books.

Phillips-Wren, Gloria. 2012. AI tools in decision making support systems: A review. *International Journal on Artificial Intelligence Tools* 21: 1240005.

Raji, Inioluwa, and Joy Buolamwini. 2019. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. *In Proceedings of the 2019 Conference on Artificial Intelligence, Ethics, and Society (AIES 2019*). AAAI/ACM.

Ryan, Tracii, Andrea Chester, John Reece, and Sophia Xenos. 2014. The uses and abuses of Facebook: A review of Facebook addiction. *Journal of Behavioral Addictions*. https://doi.org/10.1556/JBA.3.2014.016.

Sanders, Adam, Chola Elangeswaran, and Jens Wulfsberg. 2016. Industry 4.0 implies lean manufacturing: Research activities in industry 4.0 function as enablers for lean manufacturing. *Journal of Industrial Engineering and Management* 9: 811–833.

Scheutz, Matthias. 2002. Computationalism: The next generation. In *Computationalism: New Directions*, pp. 1–21.

Schwab, Klaus. 2016. *The Fourth Industrial Revolution*. Geneva: World Economic Forum.

Tabaka, Marla. 2017. Here's what's possibly causing your smartphone separation anxiety. *Wired*. https://www.inc.com/marla-tabaka/brain-hacking-why-you-have-smartphone-separation-anxiety.html

Vallor, Shannon. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting*. New York: Oxford.

Vallor, Shannon, and G.A. Bekey. 2017. Artificial intelligence and the ethics of self-learning robots. In *Robot ethics 2.0*, ed. P. Lin, L. Abney, and R. Jenkins, 338–353. Oxford: Oxford University Press.

Vasconcelos, Marisa, Carlos Cardonha, and Bernardo Gonçalves. 2018. Modeling epistemological principles for bias mitigation in AI systems: An illustration in hiring decisions. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 323–329.

Wakefield, Jane. (2016). BBC News. Microsoft chatbot is taught to swear on Twitter. Retrieved April 12, 2018, from http://www.bbc.co.uk/news/technology-35890188

Westerman, George, Didier Bonnet, and Andrew McAfee. 2014. The nine elements of digital transformation. *MIT Sloan Management Review* 55: 1–6.

Widyanto, Laura, and Mark Griffiths. 2006. 'Internet addiction': A critical review. *International Journal of Mental Health and Addiction.* https://doi.org/10.1007/s11469-006-9009-9.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.