AI and Ethics (2022) 2:1–4
https://doi.org/10.1007/s43681-021-00075-y

OPINION PAPER

The ethical AI—paradox: why better technology needs more and not less human responsibility

David De Cremer[1] · Garry Kasparov[2]

Received: 13 June 2021 / Accepted: 17 June 2021 / Published online: 24 June 2021

Abstract

Because AI is gradually moving into the position of decision-maker in business and organizations, its influence is increasingly impacting the outcomes and interests of the human end-user. As a result, scholars and practitioners alike have become worried about the ethical implications of decisions made where AI is involved. In approaching the issue of AI ethics, it is becoming increasingly clear that society and the business world—under the influence of the big technology companies—are accepting the narrative that AI has its own ethical compass, or, in other words, that AI can decide itself to do bad or good. We argue that this is not the case. We discuss and demonstrate that AI in itself has no ethics and that good or bad decisions by algorithms are caused by human choices made at an earlier stage. For this reason, we argue that even though technology is quickly becoming better and more sophisticated a need exists to simultaneously train humans even better in shaping their ethical compass and awareness.

Keywords AI ethics · Mirror · Paradox · Decisions versus choice · Behavioral business ethics

There is no doubt that AI has become part of the business world and is here to stay. The potential of AI in terms of economic benefits is unrivalled. This emerging intelligent technology is even considered by many to be more important and impactful than the internet was [1]. It is then also no surprise that AI is increasingly involved in decision-making, either as a tool, advisor or even manager [2]. This means that today intelligent technology is increasingly acquiring power to influence a wide variety of outcomes important to society. As we all know, with greater power also comes greater responsibility. For this reason, we need to start addressing the question of whether AI is intrinsically equipped to be a responsible actor and as such act in ways that we humans—as the important end-user—consider ethical.

This question is receiving much attention as the adoption of AI has created ethical concerns about, among others, privacy (compromising personal information), biased decisions (based on flawed historical data), lack of transparency (how decisions are made), and the risk to lose one's job due to automation. With such ethical concerns, fear and even anxiety about the employment and advancement of AI has surfaced in society and business. Interestingly, the narrative that surrounds the discussion about the ethicality of AI is characterized by the tendency to attribute human-like qualities to AI [3]. Because of this tendency—referred to as anthropomorphism—we seem to create the impression that AI itself can be inherently bad or good. As we tend to attribute AI such magical and human-like powers, a trend is emerging to see this intelligent (and thus learning) technology as the one being responsible for its actions and decisions. What can we learn from this trend?

This perspective identifies the important role that humans' expectations about a machine plays. Specifically, a kind of illusion seems to be in play where our enthusiasm for the supposedly magical powers of AI has led us down a road in which we essentially reduce ethics to a technological issue. How? First of all, developments in computer science contribute to this kind of thinking as fairness and ethics in this field is increasingly being seen as the same as transparency and intelligibility. Both features can be optimized by modifying technological features to algorithmic solutions [4]. Second, the developments taking place in the big tech industry also adopt a narrative that introduces ethics as a

✉ David De Cremer
bizddc@nus.edu.sg

[1] Centre On AI Technology for Humankind (AiTH), NUS Business School, National University of Singapore, 15 Kent Ridge Drive, Singapore 119245, Singapore

[2] Renew Democracy Initiative (RDI), New York, NY, USA

technological solution. Take, for example, Google's ethics-as-a-service message, which conveys to business leaders the idea that ethics is something that can easily be fixed if you have the right technology at hand. For some, this kind of message is typical of the Silicon Valley attitude, which finds solutions for almost all problems in technology. Recall Mark Zuckerberg's response in the 2018 Congressional hearings where he responded to questions from lawmakers in the House and Senate by referring to the unlimited power of AI to fix all kind of problems ranging from hate speech, fake accounts, racially discriminatory ads, to terrorist content and recruitment. As a result, we see that the practice of being ethical in the business world is transforming gradually more into an issue of technical competencies. Rather than wondering whether they should still pay attention to ethics themselves, business leaders are starting to think: "ethics, isn't that what we have AI for now?"

But, if we are honest, reality is that there is nothing magical about AI. AI does not say or demonstrate anything new that does not exist yet in the data that it is learning from. As such, AI cannot be considered to be an entity that has its own intentions, which are necessary to initiate and decide in autonomous ways to show good or bad behaviour. True, AI can move across the line between good and bad, but can only do so as a function of the intentions of the human or organization that is employing this technology. Our main point is thus that AI itself does not decide to do good or bad for the simple reason that it has no ethics. Even more so, it is our opinion that AI cannot be made intrinsically more ethical than any other technology (or even humans) just because it is "intelligent". The reason for this is that AI is a technology that is built by humans and therefore basically acts as a *mirror* to our biases. Consider, for example, the recent saga in the UK where algorithms were used to predict the results of A-Levels students based on how the secondary schools have scored historically. This algorithm-driven approach revealed an unethical outcome as many students' results were downgraded, particularly those from poorer schools. What happened is that the use of algorithms, meant to reduce teachers' bias in predicting the students' results, in reality amplified this bias. So, the reality is that complaining about bias in AI is like complaining about the image in the mirror. Hence, because the "mirror" image of AI exposes biases and flaws in our human thinking, we cannot expect AI to be suddenly more ethical than us. This also means that we have to leave behind the idea that we can trivially design machines that are more ethical than we are in the same way a programmer can create a chess program that is far better at chess than they are. We cannot!

What does our view teach us on how to manage and employ AI at work and society? We agree that AI can be used to optimise informational trends in data to create more transparency and accuracy in terms of the predictions we make. And, from this point of view, we regard AI acting as a mirror of our own biases to be a useful and powerful learning tool. Indeed, being confronted with an accurate mirror of how easily we display biased behaviour should make us appreciate AI as a tool that can help to identify our biases—of which we may not be aware—and learn from it to eliminate them where possible. Such an approach, however, does make it clear at the same time that to diminish—or even eliminate—the influence of biases on our behaviour in the future remains essentially a human responsibility. As such, we cannot pretend that it is possible to pass human responsibility and accountability over to algorithms. The phrase "the algorithm did it" should not even be part of our vocabulary. Nevertheless, we seem on the verge to consider such use of language to be legitimate.

Take, for example, the recent report from the UN Security Council revealing that an autonomous drone attacked humans without being specifically ordered to [5]. This news item gathered much attention and soon opinions converged that drones had become autonomous war tools that required no human controller anymore. Such perspective implies that the actions of an autonomous algorithm are considered separate from human decision-making. If this is truly the case, then this perspective makes it legitimate for us to look at algorithms as responsible for the possible unethical actions it will undertake. According to us, however, this idea is tantamount to saying that a gun fired itself after a person pulled the trigger. Indeed, what people seem to forget is that these drones at an earlier stage in time were programmed to attack targets and that coordinates were loaded into the software by humans. So, even though the drone autonomously decided on the attack (based on the relevant statistics), it cannot be seen as separate from human decision-making. It is thus difficult to say that the algorithm did it, because the action of hunting down a person by a drone was still the result of the actions that humans made beforehand. The work of the German and American computer scientist Joseph Weizenbaum is especially relevant here. In his 1976 book "Computer power and human reason: From judgment to calculation", he notes that computers can decide on an action that has been programmed, but it cannot choose. The reason for this is that making a choice is a product of judgment, and only humans possess the ability to make a judgment call. So, using Weizenbaum's narrative, the drone decided to attack, but it was ultimately the human that chose to upload the relevant information in the code which resulted in the launch of the drone.

All of this makes it clear that AI can show bad or good behaviour, but in its current capacity cannot be considered as the one responsible for the display of those behaviours. Whether bad or good behaviour will follow from the use of AI will be and remains a human responsibility. This viewpoint has clear implications for what AI ethics is

really about. And, to be clear, it is not the Silicon Valley narrative of big tech companies advocating that ethics has become a technology issue and that we best leave people out of it. Such a mindset can be regarded as a convenient way to facilitate the possibility for those companies to blame anything else except themselves when something goes wrong. For example, consider Facebook's Yann LeCun suggesting that as biased data lead to biased algorithms, it should follow that one should find de-biased data and let the algorithm do its work [6]. And, if the so-called indifferent algorithm—unable to be biased—still does not work then Google will come over and fix the tech [7]. Clearly, big tech companies state that ethics is important, but as many analyses have shown by now is that at the end of the day the reality for these companies is that technology is not going to wait for us (see Mark Zuckerberg's claim during the Cambridge Analytica case that his main responsibility is to innovate and provide the best technology possible). For this reason, speed and creating an environment that enables technology to innovate is what matters most making that ethics cannot be allowed to disrupt the market-dominated capitalist system that big tech company's build their reputation on (see Google's poor handling of the business-ethics balance) [8] No doubt that such a mindset indicates that the big tech companies unfortunately seem unable to take a sufficiently broad enough view of what exactly their responsibility is when it comes down to the technology that they develop.

The broader implications of the above for all of us are also clear. With the arrival and application of AI in our organizations and society, governments around the world have emphasized the need for everyone to engage in digital upskilling. A problem that we see is that in this legitimate search for more digital savviness, we seem to forget the importance to foster and even promote further our own unique human abilities that machines does not have. In other words, we fear that with an almost obsessive focus on digital upskilling, we're also creating a situation where humans will pay less attention to their strengths and as a result may lose their unique social skill powers over time. Such an outcome will only serve to harm society in the long term. And, this will especially be the case when it comes down to the ethical and responsible use of AI. As we've indicated, in the case of AI ethics, it is clear that how intelligent technology is used can only intentionally and intrinsically be determined by humans and their own ethical compass and awareness. For that reason, we suggest that scholars and practitioners alike, should be encouraged in creating more awareness that intelligent technology as it exists today cannot be a substitute for a human ethical compass. Instead, in addition to enhancing technological features that can help make data analyses more transparent and thus also more interpretable, we need to have human decision-makers that are especially more educated in ethics. Specifically, human decision-makers will need to be trained even more than ever to think through the ethical implications of decisions and be more aware of the ethical dilemmas out there.

An important conclusion is therefore that in addition to digital upskilling—which today is encouraged globally—we will need to invest more in human upskilling, and especially so in the field of ethics. We need to become better skilled at understanding our own good and bad behaviour and apply those insights to interventions and training sessions on how to use intelligent technologies in more responsible ways. Such awareness training of what we call the psychological underpinnings of (un)ethical behaviour can teach us when humans are most likely to show unethical behaviour and translate those into the settings of designing and employing intelligent technology [9]. As such, the development of ethical AI will have to be founded on an interdisciplinary approach between computer science and social sciences to arrive at an understanding that will enable humans to use intelligent technology to—at the same time—augment their abilities while being able to make decisions in efficient yet ethical ways.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. CEOs: AI will have larger impact than the internet. https://www.marketingcharts.com/customer-centric/analytics-automated-and-martech-107328 (2019). Accessed 1 June 2021
2. Haesevoets, T., De Cremer, D., Dierck, K., Van Hiel, A.: Human-machine collaboration in managerial decision making. Comput. Hum. Behav. (2021). https://doi.org/10.1016/j.chb.2021.106730
3. De Cremer, D., Kasparov, G.: AI should augment human intelligence, not replace it. Harvard business review. https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it (2021). Accessed 1 June 2021
4. De Cremer, D.: What does building a fair AI really entail? Harvard business review. https://hbr.org/2020/09/what-does-building-a-fair-ai-really-entail (2020). Accessed 1 June 2021
5. Sankaran, V.: Military drones may have attacked humans for first time without being instructed to, UN report says. Independent. https://www.independent.co.uk/life-style/gadgets-and-tech/drone-fully-automated-military-kill-b1856815.html (2021). Accessed 1 June 2021
6. Machines are indifferent, we are not: Yann LeCun's Tweet sparks ML bias debate. https://analyticsindiamag.com/yann-lecun-machine-learning-bias-debate/ (2020). Accessed 1 June 2021
7. Simonite, T.: Google offers to help others with the tricky ethics of AI. https://www.wired.com/story/google-help-others-tricky-ethics-ai/ (2020). Accessed 1 June 2021
8. Grant, N., Bass, D., Eidelson, J.: Google Turmoil exposes cracks long in making for top AI watchdog. Bloomberg. https://www.bloomberg.com/news/articles/2021-04-21/google-ethical-ai-group-s-turmo

il-began-long-before-public-unraveling (2021). Accessed 1 June 2021
9. De Cremer, D., Moore, C.: Toward a better understanding of behavioral ethics in the workplace. Annu. Rev. Organ. Psych. Organ. Behav. **7**, 369–393 (2020)