AI and Ethics (2022) 2:579–583
https://doi.org/10.1007/s43681-021-00116-6

OPINION PAPER

The road to a human-centred digital society: opportunities, challenges and responsibilities for humans in the age of machines

David De Cremer[1] · Devesh Narayanan[1] · Andreas Deppeler[1] · Mahak Nagpal[1] · Jack McGuire[1]

Received: 21 October 2021 / Accepted: 27 October 2021 / Published online: 16 November 2021

Abstract

The growing adoption of intelligent technologies has brought us to a crossroad. The creators of intelligent technologies are acquiring the power to influence a wide variety of outcomes that are important to human end-users. In doing so, those same intelligent technologies are being used to undermine and even actively harm the interests of those same end-users. In the absence of a recalibration, we are almost certainly headed down a path wherein intelligent technologies will primarily serve the interests of developers and owners of technology rather than humankind at large. In an attempt to push for such a recalibration, we present parallels between the 2008 financial crisis and the current state of affairs. Following which, we present a list of recommendations and implications to be used when in the pursuit of creating responsible and human-centred AI.

Keywords Human-centred AI · AI ethics · Tech crisis · Techno-solutionism · Manifesto · Big tech companies

Intelligent technologies are dramatically transforming modern societies. The potential economic and social benefits of these technologies seem unprecedented. Intelligent technologies are, therefore, increasingly being involved in a variety of decision-making contexts: to support, advise and sometimes even override human decision-makers [1, 2]. As a result, as organizations undergo digital transformations, such technologies are increasingly used to influence a wide variety of outcomes that are important to human end-users [3]. But of course, with greater power also comes greater responsibility. As such, it is no surprise that a strong need is emerging for greater scrutiny about the extent to which humans are vulnerable to the actions and decisions of intelligent technologies.

These vulnerabilities materialize in various ways. For example, intelligent technologies can undermine or sometimes even actively harm the interests of their human end-users, leading to unfavourable or unethical outcomes. Quantification and datafication are crucial to the functioning of AI but may be perceived by users as depersonalizing and reductionist [4, 5]. More generally, as intelligent technologies become part of our lives, the risks to the stability of our social fabric and the sanctity of our human autonomy are becoming increasingly apparent. Growing concerns about these risks and harms threaten to undermine many of the benefits that these technologies could create.

At the Centre on AI Technology for Humankind's (AiTH), its founder and director David De Cremer has previously noted, technology will and can definitely be used for good [6]. However, current investments in intelligent technologies and automation are largely driven by cost-cutting motives: lured by the prospects of growth without having to raise salaries or hire more people. If these cost-cutting efforts are not combined with investments in what we call "human upskilling"—where the abilities, actions and interests of humans are cultivated and refined with the support and assistance of technology—then we fear that the harms and vulnerabilities listed earlier will surely materialize. The obsessive search for technological solutions striving to optimize efficiency and maximize productivity will prioritize investments in innovations that primarily serve the interests of those designing and distributing intelligent technologies. Following such a path will lead to building a technologically regulated society that serves the interests of machines and their developers, rather than humankind at large.

Even more so, such a society would be drastically different from the one imagined in the past. For example, the British economist John Maynard Keynes predicted

✉ David De Cremer
bizddc@nus.edu.sg

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

almost a century ago that technical innovation would improve labour productivity and overall wealth to such an extent that working 3 h a day would be "quite enough" [7]. Today, however, people are working more than ever (further exacerbated by the COVID-19 pandemic, which made working from home the new default), salaries in many professions and regions have been stagnant (in real terms) since the 1980s, and pension funds are under threat everywhere—forcing people to work longer than before. At the same time, the wealth created by the increasing use of technology is accumulating in the coffers of a few large and powerful technology companies—and in the private accounts of the largest shareholders. Take, for example, the message of Jeff Bezos after he took a short trip to "space" in his Blue Origin spacecraft. Upon his return to earth, he thanked the Amazon employees and customers, because they paid for his trip. Not surprisingly, people could not appreciate his message [8, 9].

Big Tech companies employ a certain strategy accompanied by a specific narrative. This is the narrative of "techno-solutionism"—that technology can be used to solve most problems that we encounter in society and business. This typical Silicon Valley mindset and narrative has permeated governments and businesses, entrenching beliefs that social problems can be 'solved' if one has the right technology. Indeed, because of this techno-solutionist mindset, we have come to see most societal problems and challenges as ones that can easily be optimized by modifying the properties of a machine-learning algorithm. Take, for example, Google's recent announcement of "ethics-as-a-service", which conveys to business leaders the idea that if algorithmically made decisions are unfair or biased, this can be 'fixed' with certain technical tweaks to those same algorithms and datasets [10]. As another example, call centres have been recently emphasizing that their employees need to be trained to act in more empathic ways to their customers [11]. Tech start-ups, in turn, have offered a solution: to use algorithms, that have been trained to imitate and recognize empathy, to coach the call-centre employees to be more empathetic. This is a strange logic: this problem only exists, because the call-centre model incentivizes workers to be less empathetic, whilst incentivizing mechanical, output-oriented behaviours. In other words, this problem only exists because of the inefficiencies and value-misalignments embedded in the technology, but the solution offered to address it is once again technological.

What do the above examples illustrate? To us, they show how intelligent technologies—and the companies that design and develop them—have acquired a position of power that apparently goes unchallenged. Technology is the only way forward, and if its use reveals problems, the solution involves more technology. It sounds like a path is being paved for a world that is more suited for machines than for humans.

At AiTH, we are deeply concerned about these seemingly "machine-centred" approaches to the design and deployment of AI. The adoption of reductionist perspectives—where finding the right incentives and rewards is deemed sufficient to optimize for the behaviours and decisions we want to see—is a profound threat to our humanity. Machine-centred approaches threaten to box in our complex, authentic social lives, in their bid to reduce humans to quantifiable and predictable data objects. In contrast, a human-centred approach to developing and deploying intelligent technologies appreciates the complexities and grey zones where human judgments and intuition will always be needed, and strives to harness the potential of AI to serve the needs of, and create benefits for, humans.

1 Repeating history?

Intelligent technologies are popularly construed as a powerful exogenous force—sweeping in to disrupt our social and work lives at a rate that most humans cannot keep pace with. As AiTH director David De Cremer has previously noted, the magical thinking surrounding AI has caused many businesspeople to worry about finding a place for humans in a world run by computers, rather than the other way around [3]. Such thinking, and the fear of humans being "left behind", threaten to fragment our social fabric. Perceived divides—between organizations that are "AI leaders" and those that are "AI laggards", between those whose jobs will be "disrupted" and those whose jobs are "safe", and between "technophiles" and "luddites", to name a few—can produce widespread social anxiety and dissatisfaction among those who feel left out of the technological future we seem to be hurtling towards.

Continuing on a path of accelerating technological developments without reflection and critical analysis based on a human-centred perspective will, in our view, lead to a new economic and social crisis. This upcoming 'tech crisis' would draw upon and reinforce the various social anxieties and fears about intelligent technologies that we are already starting to see. However, like all previous crises in recent history, its causes and effects on society and the economy are likely to be much broader. Indeed, we believe that there are useful and productive parallels to be drawn between the potential tech crisis to come and the global financial crisis of 2007–2008.

For one, the mindsets shared in the corporate world seem to be surprisingly similar between then and now. The promotion of calculative and hyper-competitive thinking and a 'ticking-the-box' mentality seem to be characteristic of contemporary technoscientific thinking where reducing everything (including humans; cf. people analytics) to measurable and predictable data-points is omnipresent. These same

practices and mindsets were also dominant in the wake of the global financial crisis. Indeed, although financial engineering, and overconfidence in mathematical models and quantitative risk management were not the proximate cause of the global financial crisis, they provided a false sense of security. Mortgage origination and trading desks at major investment banks relied on quantitative models and ever more complex financial engineering to manage the risks of their holdings. It gave the impression—to senior management, regulators, investors, rating agencies and the general public—that risk could be quantified and controlled. An overreliance on models, quantification and rationalization clouded people's judgments and prevented them from recognizing the looming dangers until it was too late. We have since learned that calculative and hyper-competitive thinking drives unethical behaviour and the tendency to justify ambiguous ethical decisions [12, 13], so it is especially worrying to notice the ascendancy of such thinking again in today's tech-dominated corporate world.

Furthermore, in the wake of the global financial crisis, banks were seen as "too big to fail" because they own and run most of the financial infrastructure that is essential to the functioning of the globalized economy. Similar beliefs seem to be held about tech companies today. A handful of tech companies provide and maintain the digital cloud infrastructure that supports much of the world's private and public sector activities. They are, so to speak, too essential to the modern digital economy to fail [14]. Of course, we now know that such ways of thinking lull us into a false sense of certainty which leads to complacency. These institutions are glued together by long, and ultimately fragile chains of trust. When people—and, in particular, veto players such as powerful monied interests and governments—lose trust, the whole system collapses. We believe that such a collapse of trust in the tech-dominated corporate world is be possible.

The belief in technocratic solutions is part of a broader ideology of instrumental rationality. We saw it in the early 2000s when banks relied on oversimplified models to manage complex structured financial products. We have seen it since the crash when central bankers devised ever more creative ways to engage in what essentially amounts to printing money. Today, we see it in pronouncements of technology firms about ethical and responsible artificial intelligence. It is the unquestioning belief in the inevitability of technical progress, along with the assumption that any potential threats or harms arising from such never-ending progress can be "managed" or "mitigated" with ever more technical solutions. The result is an epistemic bubble that enwraps all public discourse. Piercing the bubble requires humility, clarity of mind and a historical perspective. Unfortunately, these traits are all too rare among today's leaders in governments and corporations and for that reason poses a threat to how we design and implement regulatory and policy instruments.

Indeed, one final analogue that we observe is the one encapsulating the relationship between the corporate world and regulators. The business world, both in the past and today, looks to regulators as *suppliers* of norms. Regulators, on the other hand, are often unwilling or unable to enforce strict regulations on businesses—usually citing concerns about the adverse effects of blunt regulation on productivity and competitiveness. Instead, a culture of self-regulation is emphasized. Companies are the ones who are supposed to ensure that their actions are ethical, while regulators may, at best, provide some frameworks and guidance to help with this self-regulation. We know now that such self-regulation was insufficient to stop the devastation of the global financial crisis. Despite this, we see it being pursued once again in the tech industry. The past few years have seen numerous frameworks, principles and policy documents, but little in the way of actually enforceable regulation [15, 16]. High-level ethical standards prove difficult to translate into actual behavioural actions. We do have doubts that such frameworks will serve any meaningful role in stopping the impending tech crisis.

2 A manifesto for responsible and human-centred AI

To document our fears and concerns that currently surround the development of AI technologies, which may lead to a potential "tech crisis', AiTH decided to write the manifesto that you are reading now. Specifically, the aim of the manifesto is provide a public opinion about the state of discourse on AI ethics and trustworthiness, the unquestioned dominance of Big Tech, and the deficiencies of techno-solutionist and machine-centred approaches to AI. In particular, with this manifesto, we wish to make clear what we believe is needed (and why) to employ a distinctive and legitimate human-centred approach to the adoption and integration of intelligent technologies in our businesses and society.

2.1 AiTH's approach to human-centred AI

A Human-Centred approach to AI (HCAI) focuses on designing and deploying AI systems in ways that serve the needs of, and create benefits for, humans. In line with this purpose, we recognize that HCAI must contribute to and empower human's experience of competence, sense of belonging, control and well-being.

1. Competence:
   HCAI augments and enriches human capabilities and performance across all domains in life, rather than automating away the skills and attributes that make us human.

2. Belonging:
   HCAI designs AI systems with the understanding that intelligent technologies are fully embedded in society. Such systems can, therefore, be expected to act in line with the norms and values of a humane society, including fairness, justice, ethics, responsibility and trustworthiness.
3. Control:
   HCAI preserves human agency and sense of responsibility by designing AI systems to give users a high level of understanding of, and control over, their specific and unique processes and outputs.
4. Well-being:
   HCAI advances the self-esteem, confidence and happiness of all humans. The design and deployment of such AI systems must be mindful to the varied dimensions of life that they stand to impact, as well as their long-term effects on overall well-being.

2.2 Implications and recommendations

Our recommendations for businesses and policymakers derive directly from AiTH's research and thought leadership based on the four facets of HCAI. The following high-level recommendations aim to provide guidance on the types of considerations that need to be made while pursuing human-centred AI.

1. Humans first, machines second:
   The capabilities of intelligent technologies for thought and action should not serve as the standard by which humans are assessed and compared. Considerations about the well-being and flourishing of humans—especially those systematically disadvantaged and disenfranchised—must always be central to any technology deployment. *We should be preparing machines to serve humans, rather than preparing humans to serve machines.*
2. 'Digital transformation' and the adoption of intelligent technologies should be value-driven rather than solely profit-driven:
   We can use machines for good if we are clear about what our human identity is and what value we want to create for a humane society. A clear understanding of how to do business and what kind of value ought to be created for end-users can serve as a lens for evaluating the appropriateness and necessity of technological interventions.
3. Human and machine intelligences should not be treated as interchangeable:
   Automation should not be thought of in terms of its potential to replace or disrupt human labour. Instead, in line with the 'augmentation' paradigm, we should instead evaluate automation in terms of how it complements and enhances our human abilities and ways of working. The future of work should be a collaborative one: where machines are deployed in ways that respect the autonomy and abilities of workers, and in turn, make work better for everyone.
4. The ultimate responsibility for technological-augmented decisions must remain in human hands:
   Intelligent technologies are not moral agents. The 'decisions' they make are situated within contexts and rules set in place by human choices—by those who develop, deploy and use them. As such, humans must and are obliged to retain ultimate responsibility for these decisions.
5. Ethical considerations about technology must be embedded in organizational structures and practices, rather than in abstract frameworks and principles:
   Current governance frameworks and principles reduce ethics, fairness and trust to technological features and boxes to be ticked. However, we can only have 'ethical AI' when ethics is fully integrated into daily organizational life. Leaders need to translate principles into specific practices, and moral upskilling is needed for all workers.
6. Embrace value pluralism and respect cultural differences while advancing ethical AI:
   Current conversations about ethical AI tend to emphasize perspectives from the West rather than the East, and the Global North rather than the Global South. This trend is at least partly due to greater attention to the topic in these regions, which necessitates a call for more thought leadership in this field to emerge from Eastern counterparts (a role that AiTH proudly takes upon itself). For human-centred AI to serve the needs of *all*—rather than just a few—humans, we must be sensitive to how values and interests are displayed differently across diverse cultural and social contexts and how these differences may impact our thinking about and assessment of fair, trustworthy and ethical intelligent technologies.
7. Focus on real AI, rather than imagined AI:
   There are growing tendencies to focus on the anticipated future risks and benefits of certain kinds of 'superintelligent' AI that might exist in the future. Such imaginaries tend to distract and obfuscate the real harms and benefits that 'narrow' AI systems are already bringing to organizations and society. Fantasies of "superhuman" AI (endlessly repeated by business writers and self-proclaimed experts) mislead people into overestimating the capabilities of currently available AI systems. As a result, today's society runs the risk to be constructed and shaped in correspondence with imaginaries of AI—a world more suited for machines—that may or may not materialize. We hasten to say that we do not mean to

suggest that superintelligence is strictly impossible. However, the process of building value-aligned and human-centred AI must begin with a realistic attitude that focuses on the AI systems that we have today, and the actual material harms and benefits that they presently create.

Declarations

Conflict of interest The authors declare that they have no conflicting interests.

References

1. De Cremer, D.: With AI entering organizations, responsible leadership may slip! AI Ethics (2021). https://doi.org/10.1007/s43681-021-00094-9
2. Chamorro-Premuzic, T., Wade, M., Jordan, J.: 'As AI Makes More Decisions, the Nature of Leadership Will Change', *Harvard Business Review*, Jan. 22, 2018. Accessed: Oct. 21, 2021. [Online]. Available: https://hbr.org/2018/01/as-ai-makes-more-decisions-the-nature-of-leadership-will-change
3. De Cremer, D.: Leadership by algorithm: who leads and who follows in the AI era? Harriman House Limited, Petersfield (2020)
4. Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions. arXiv (2018). https://doi.org/10.31235/osf.io/9wqxr
5. Newman, D.T., Fast, N.J., Harmon, D.J.: When eliminating bias isn't fair: algorithmic reductionism and procedural justice in human resource decisions. Organ. Behav. Hum. Decis. Process. **160**, 149–167 (2020)
6. World Economic Forum, 'Machine learning or leading? This is how AI will influence leadership according to an expert', *World Economic Forum*, 2020. Accessed: Oct. 21, 2021. [Online]. Available: https://www.weforum.org/agenda/2020/11/artificial-intelligence-ai-algorithm-leadership/
7. Keynes, J.M.: Economic possibilities for our grandchildren. In: Essays in persuasion, pp. 321–332. Springer, Berlin (2010)
8. Goodkind, N.: 'Bezos says Amazon workers and customers paid for his trip', *Fortune*, 2021. https://fortune.com/2021/07/20/jeff-bezos-thanks-amazon-workers-and-customers-after-space-flight-you-paid-for-all-of-this/ (Accessed Oct. 21, 2021).
9. Marx, P.: 'Amazon workers don't want a Bezos shoutout. They would like a raise, though.', *NBC News*, 2021. https://www.nbcnews.com/think/opinion/jeff-bezos-thanks-amazon-workers-blue-origin-launch-revealingly-tone-ncna1274565 (Accessed Oct. 21, 2021).
10. Simonite, T.: 'Google Offers to Help Others With the Tricky Ethics of AI', *Wired*, 2020. Accessed: Jul. 27, 2021. [Online]. Available: https://www.wired.com/story/google-help-others-tricky-ethics-ai/
11. Yue, F.: 'AI at work: Machines are training human workers to be more compassionate', *USA TODAY*, 2019. https://www.usatoday.com/story/tech/2019/08/23/ai-training-human-employees-to-have-more-empathy-work/2070002001/ (Accessed Oct. 21, 2021).
12. Wang, L., Zhong, C.-B., Murnighan, J.K.: The social and ethical consequences of a calculative mindset. Organ. Behav. Hum. Decis. Process. **125**(1), 39–49 (2014)
13. Pierce, J.R., Kilduff, G.J., Galinsky, A.D., Sivanathan, N.: From glue to gasoline: How competition turns perspective takers unethical. Psychol. Sci. **24**(10), 1986–1994 (2013)
14. Withers, I., Jones, H.: For bank regulators, tech giants are now too big to fail, *Reuters*, Aug. 20, 2021. Accessed: Oct. 21, 2021. [Online]. Available: https://www.reuters.com/world/the-great-reboot/bank-regulators-tech-giants-are-now-too-big-fail-2021-08-20/
15. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
16. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020)