AI & SOCIETY (2022) 37:1319–1322
https://doi.org/10.1007/s00146-022-01582-0

EDITORIAL

Machine learning: can the automatic pilot transcend the toxic fog?

Karamjit S. Gill[1]

Accepted: 17 October 2022 / Published online: 31 October 2022

Whilst the media is intrigued by the humanoid robot Ai-Da addressing the British House of Lords committee on the future of the arts, design, fashion and music industries, telling a committee that artificial intelligence can be a 'threat and opportunity' to artists (The Guardian 2022), we also note concerns of the media on the risk of creating a generation of racist and sexist robots'. Andrew Hundt, of Georgia Tech warns that 'we're at risk of creating a generation of racist and sexist robots', and of artificial intelligence becoming bigoted after learning 'toxic stereotypes' on the internet'. He says that 'The robot has learned toxic stereotypes through these flawed neural network models,' and that it is not ok for 'people and organisations to create these products without addressing the issues.' (The Daily Mail 2022). Selinger (2019) notes that although the contributors recognise Wiener's concerns (Norbert Wiener, The Human Use of Human Being, 1950), such as those of cultural anxiety of automation, nature of surveillance, and social risks stemming from careless integration of machine-generated decisions with governance processes and misuse (by humans) of such automated decision making, Selinger argues that postponing work on ethical issues until after goal-aligned AGI is built would be irresponsible and potentially disastrous. He further says that perfectly obedient superintelligence whose goals automatically align with those of its human owner would be like 'Nazi SS-Obersturmbannführer Adolf Eichmann' on steroids: lacking a moral compass or inhibitions of its own, it would, with ruthless efficiency, implement its owner's goals, whatever they might be. Extending the debate on the tension between ruthless efficiency and ethical constraints, Shauna Concannon (2021) discusses the role of virtual personal assistants and other forms of `conversational AI' in health and social care and asks whether an AI system can perform caring duties or offer companionship. Virtual personal assistants such as Siri, and Alexa are designed to respond to users in ways that create the illusion that they understand something of the user's psychological or emotional state. However, empathy is often thought of as a uniquely human trait that enables us to form connections with and understand one another. As chatbots designed to support wellbeing and perform therapeutic functions are already available and widely used, a question arises: could or even should they be able to empathise with their users, and further what are the ethical implications that arise when positioning AI systems in roles that require them to communicate with empathy?

In this volume, we look back at the contribution of our authors and their reflections on these tensions arising from unregulated system such as discriminatory facial recognition and predictive AI systems and policing strategies. These discriminatory systems pose social challenges of governance, ethics, accountability and intervention arising from the accelerated integration of powerful artificial intelligence systems into core social institutions. Helga Nowotny (2021) proposes that there is a tacit assumption and misplaced confidence that ethical AI would ultimately take care of the unresolved ethical, transparency and accountability conflicts when we are able to develop computational tools 'to assess the performance and output quality of Deep Learning algorithms and to optimise their training'. The danger, she says, is that we end up trusting the automatic pilot while flying blindly in the fog, becoming part of a fine-tuned and interconnected predictive system, thereby diminishing our motivation and ability to stretch the boundaries of imagination.

But what drives this idea of the ethical machine? First, the desire to seek objectified solutions without prejudice in the scientific tradition; second, belief in calculation as measurement of objectification; third, confusion in the idea that data is objectivity and not calculation; fourthly the idea of machine ethics as an extension of human ethics, ultimately becoming fully aligned with the machine's operations—just as the machine was seen as an extension of the human body, now machine intelligence is seen as an extension of human intelligence. Those who are engaged in the pursuit of machine ethics and governance are reminded that actionable ethics is also about the pursuit of inclusive participation

✉ Karamjit S. Gill
editoraisoc@yahoo.co.uk

[1] University of Brighton, Brighton, UK

and openness towards knowledge of the past, complexities of the present and uncertainties of the future. In the end, it is not important how the AI machine can be aligned with human values or visualising as how human values are fully aligned with the AI machine, converging to the post-human world, what it is important to know is that human values are diverse, social, cultural and contextual, and they do not fit into the logic of the AI machine.

In this volume, we look back on reflections of our authors on the debate on the tension between functional efficiency and ethics form multiple perspectives. AIS Vol. 37.1 on 'Actionable Ethics' covers a range of issues: framing AI systems in healthcare sector; social machine as a tool for shaping interactions between individuals and algorithms; algorithmic accountability, transparency and intentional biases; algorithmic augmentation of democratic processes, discrimination in the age of artificial intelligence; ethics and biometric facial recognition technology; algorithmic and human decision making, and standards of transparency; explainable artificial intelligence and its intrinsically value and desirability; how do people judge the credibility of algorithmic sources; shifting relations of human autonomy and technological automation; ethical challenges and organisational responses on responsibilities of policymakers, professional bodies and regulators; data objects for knowing-data science a technology-driven science; endowing artificial intelligence with legal subjectivity; in search of the moral status of AI; actionable ethics for governance; sensorimotor debilities in digital cultures; social acceptance of robots; child-robot relationship formation; AI machine and the art of education; the challenge of defining cross-cultural fairness assessments of texts; multifaceted nature of the transformation; impact of AI on human behaviour and emotions in a multicultural educational context; AI for seeing creativity assessment of culinary products as art; the making of AI futures in German context; the limit of human anthropocentric tendencies of control; utilitarianisms and machine ethics; Dystopian conception of posthumanism vs. Africanist civilizational humanism.

In continuing the debate on actionable ethics, the AIS Vol. 37.2 on 'Autonomous Reciprocity' explores issues of ethics, sustainability, and responsibility in social robotics. Within the context of social robots and the ethics of care, we note possible effects of nudging in reciprocal relationships between humans and robots. Furthermore, we are alerted to the danger of designing social robots for reciprocity where reciprocity may be used as an instrumental value to enhance acceptability of the robot, and this is ethically questionable. However, in contrast, we also learn how humans develop empathic responses to robots. This argument on the ethical reciprocity draws upon the philosophy of the Danish theologian K. E. Løgstrup, that human empathy is inherently good, because it turns people away from their own self-focus (inturnedness), and this concept of empathy applies also to relations with robots. Although it is acknowledged that reciprocity is indeed a component of moral development, and is in no way harmful in itself. It is, however, uncertain whether reciprocity fostered in Human Robot Interaction (HRI) would transfer to human–human interaction where it would provide the most benefit. It is thus much better to focus on fostering reciprocity among humans to facilitate human–robot interactions. We are asked to pay attention to the debate on unintended or undesirable consequences of empathic responses of human to robots, for example the potential for malicious intent and exploitation in robot design and development in the name of ethical socio-emotional relationships with robots. The notions of human–robot reciprocity and empathetic interaction highlight the oversimplification of social care and service practices in the design of human likeness in robots as social companions. The core premise of this articulation of reciprocity is that sociality is not something that can be a property of a machine, but is rather something that is enacted in an encounter, or an evolving relationship, between a human and a machine. If this is the case, then we should focus on the enactment of empathic social agency, rather than its representation, in the design of social robots.

We note how the representation of empathetic reciprocity is propagated in the design of social robotics for the care and service sectors, for example in the therapy and care of dementia patients, robot companions for older adults living at home. The idea of robots providing services that we would otherwise expect from humans forces us to think about the aspects of these services that may, and may not, be replaceable. Here, the technologies that promise remedies to human vulnerabilities seem very enticing, and this faith in technological solution of social problems leads to an oversimplification of the role of humans in care and service work, or a reduction in the complexity of the tasks that they carry out. By depicting older adults as dependent, fragile and vulnerable people, renders them as 'potentially burdensome care recipients', and robot technologies are presented as an optimal solution to this social problem. In a similar vein, the roles of caregivers and care-receivers, and care practices are deconstructed into tasks to fit well-defined technical problems. This leads to an incremental mechanization of care, rather than to a more holistic understanding of it. This oversimplification of social care, rooted in misconceptions about the provision of care, the process of ageing, affective labor in professional service work, can influence the design and implementation of social robots. Although there is a deep concern about the potential replacement of human care providers with robotic technologies, the introduction of robots as complementary technologies in social settings raises important questions of autonomy and ethics. Our attention is drawn to the debate on ethics and autonomy,

where human subjects attribute autonomy to their experience of artificial devices. The idea of autonomy, rooted in Western philosophy of Aristotle and Kant, widely assumes that any perceivable action has a 'source' that centers on an actor/agent. In some undefined sense, humans (and all living beings) are taken to act autonomously, and this view thus has consequences for living human beings. This perception of autonomy not only impinges on organizational, social and individual experiences and actions, but also on how we conceptualise AI devices such as predator drones as killer robots, and our roles as actors (and entities) and its implications for designers of such machines. The perceived autonomy is thus related to not only how autonomy is perceived but also how working with human–machine aggregates from a broader perspective of interactional and situational outcomes, socio-cognitive organization, culture and, thus, of the ethical issues that are central to AI. The question arises whether we could–indeed should build machines as moral actors, and in what ways those working in machine ethics treat the autonomy of artificial agents as quite unlike that of natural agents. If Kantian view of ethics and agency depends on the seat of reason or the mind, artificial moral agents (AMA) should not only be rational but also fundamentally subjective. From this perspective, 'Kantian AMA' would, therefore, pursue, not common interests or those of communities, but outcomes that are consistent with universal, individual and voluntarist reasoning. However, if we take Aristotelian tradition of ethics and agency in the sense that living human beings act ethically within a social context, then autonomy is not seen as intrinsic but, rather, fundamentally relational. In this case, moral judgements can only be traced to the embodied socialization of a citizen. Depending upon whether we take a Kantian view or an Aristotelian view of autonomy, the AMAs would differ in evaluating what is good and appealing, on the one hand, to society as a whole and, on the other, to a rational grasp of what is right. This rests on the view that humans, at least, exhibit the autonomy of social beings, and further depends upon how we see AMAs, how we see their societal role and, how we regulate and motivate designers.

In this volume, some of our authors reflect on designing AI systems that are concerned with Empathic AI, the Future of Consent, legitimacy of algorithmic decision systems and AI-driven social theory. Carlos Montemayor in 'In Principle Obstacles for Empathic AI' (this volume) discusses the limits of the use of Artificial Intelligence (AI) in the relational aspects of medical and nursing care and notes that many of the obstacles discussed in the literature on empathetic AI are technical in character, regarding how to improve and optimize current practices in clinical medicine and also how to develop better data bases for optimal parameter adjustments and predictive algorithms. The author notes that there are also in principle obstacles to the application of AI in clinical medicine and care where empathy is important, and that these problems cannot be solved with any of the technical tools. The technical focus is likely to generate specific risks that may be overlooked, and this necessitates human monitoring and emotional intervention in clinical medicine. In addition to the specific risks, the technical focus may raise difficult issues of moral and legal responsibility. Adam J. Andreotta et al. in 'AI, Big Data, and the Future of Consent' (this volume) discuss problems with current Big Data practices which, they claim, seriously erode the role of informed consent as it pertains to the use of personal information. To illustrate these problems, they consider how the notion of informed consent has been understood and operationalised in the ethical regulation of biomedical research (and medical practices, more broadly) and compare this with current Big Data practices. They do so by first discussing three types of problems that can impede informed consent with respect to Big Data use. First, they discuss the transparency (or explanation) problem. Second, they discuss the re-repurposed data problem. Third, they discuss the meaningful alternatives problem. In the final section of the paper, they suggest some solutions to these problems. In particular, they propose that the use of personal data for commercial and administrative objectives could be subject to a 'soft governance' ethical regulation, akin to the way that all projects involving human participants (e.g., social science projects, human medical data and tissue use) are regulated in Australia through the Human Research Ethics Committees (HRECs). They also consider alternatives to the standard consent forms, and privacy policies, that could make use of some of the latest research focussed on the usability of pictorial legal contracts. Clément Henin and Daniel Le Métayer (this volume) point out that explainability is useful but not sufficient to ensure the legitimacy of algorithmic decision systems. They argue that the key requirement for high stakes decision systems should be justifiability and contestability. They highlight the conceptual differences between explanations and justifications, provide dual definitions of justifications and contestations, and suggest different ways to operationalize justifiability and contestability. Jakob Mökander and Ralph Schroeder in 'AI and Social Theory sketch a programme for AI-driven social theory'(this volume) and lay out how AI-based models can draw on the growing availability of digital data to help test the validity of different social theories based on their predictive power. In doing so, they use the work of Randall Collins and his state breakdown model to exemplify that, already today, AI-based models can help synthesize knowledge from a variety of sources, reason about the world, and apply what is known across a wide range of problems in a systematic way. However, they also find that AI-driven social theory remains subject to a range of practical, technical, and epistemological limitations. Most critically, existing AI systems lack three essential capabilities needed to

advance social theory in ways that are cumulative, holistic, open-ended, and purposeful. These are (1) semanticization, i.e., the ability to develop and operationalize verbal concepts to represent machine-manipulable knowledge; (2) transferability, i.e., the ability to transfer what has been learned in one context to another; and (3) generativity, i.e., the ability to independently create and improve on concepts and models. They argue that if the gaps identified here are addressed by further research, there is no reason why, in the future, the most advanced programme in social theory should not be led by AI-driven cumulative advances.

References

AI&Society (2022a) Vol. 37.3 https://link.springer.com/journal/146/volumes-and-issues/37-3. Accessed 3 Oct 2022.

AI&Society (2022b) Vol. 37.1 https://link.springer.com/journal/146/volumes-and-issues/37-1. Accessed 3 Oct 2022.

AI&Society (2022c) Vol. 37.2 https://link.springer.com/journal/146/volumes-and-issues/37-2. Accessed 3 Oct 2022.

Concannon S (2021) Empathetic machines. Centre for Research in the Arts, Social Sciences and Humanities. https://www.crassh.cam.ac.uk/events/29462/. Accessed 5 Apr 2021

Nowotny H (2021) In AI we trust: power illusion and control of predictive algorithms. Polity Press, Cambridge

Selinger, Evan (2019). Why technologists fail to think of moderation as a virtue and other stories about AI. Los Angeles review of books. https://lareviewofbooks.org/article/why-technologists-fail-to-think-of-moderation-as-a-virtue-and-other-stories-about-ai/. Accessed 29 Dec 2021

The daily mail (2022) https://www.dailymail.co.uk/sciencetech/article-10957023/Fears-AI-create-sexist-bigots-test-learns-toxic-stereotypes.html. Accessed 28 Sept 2022

The guardian (2022) https://www.theguardian.com/technology/2022/oct/14/ai-da-robot-sums-up-flawed-logic-lords-debate-ai. Accessed 14 Oct2022