AI&Society: editorial volume 35.2: the trappings of AI Agency

Karamjit S. Gill^{1,2}

Published online: 12 March 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

What should be our response to AI Agency trapping us in the data-driven web of the AI-powered machine? Among wide-ranging responses of AI communities to the crisis of AI Agency, the response in general is to warn us about the existential risk posed by Super AI, the dystopia of Black Mirror, and the crisis of automated decision-making. This response makes us aware of our techno-centric future where our identities are being linked to facial recognition, and of the deep concerns about data paralysing institutions and industries, thereby creating a culture of ‘data anxiety’. The second response is to articulate the implication of opaqueness and the lack of transparency of autonomous AI systems, raising concerns of manipulation of decision-making. This response also alerts us to the impact of AI Agency on the politics of governance. The third response is to question the very nature of intelligence of the artificial, raising questions about sentience and our understanding of the data-driven world. This response also warns us of the danger of getting used to blind faith in the machine, trappings of a human–robot co-existence society, and elimination of human intervention in autonomous decision-making. Further, it alerts us of empty slogans of transparency and compliance and “ethics washing” facades. The fourth response is to counter the images of a dystopic future, asking us to give attention to positive impacts and potentials of AI systems for societal benefit in domains such as human health, transportation, service robots, health-care, education, public safety, security and entertainment. The fifth response is to initiate a conversation on public accountability frameworks, including issues of governance, and the cultivation of a culture of algorithmic accountability arising from concerns of opaqueness, transparency and responsibility. Whilst recognising the need

to cultivate trust and reliability of AI systems and tools, it argues for alignment of AI Agency to social, cultural, legal and moral values of societies, guided by ethical frameworks. To get a glimpse of varied voices and responses to the trappings of the AI Agency, we take a note of recent AI debates of forums such as those of The Economic Word Forum (2019), the STOA Study (2019), the AI and the Future of Humanity Exhibition (2020), The Royal Society (2018), and voices of the AI research community including those of the authors of this volume.

The Economic Word Forum White paper (2019) concludes that “The increasing use of AI and autonomous systems will have revolutionary effects on human society. Despite many benefits, AI and autonomous systems involve considerable risks that must be managed well to take advantage of their benefits while protecting ethical values as defined in fundamental rights and basic constitutional principles, thereby preserve a human-centric society.” It points out the lack of transparency, the increasing loss of humanity in social relationships, the loss of privacy and personal autonomy, information biases, as well as error proneness and susceptibility to manipulation of AI-powered autonomous systems. On the debate on embedding ethical principles in the AI machine, the White paper alerts us to a possible slavish adherence of AI systems to a particular ethical school of thought in decision-making. As the concern and impact of AI agency move to the politics of governance, STOA (2019) Study explores a European rationale for creating governance frameworks for algorithmic accountability and transparency. The study notes that a lack of transparency of algorithmic systems risks their meaningful scrutiny and accountability when these systems are integrated into decision-making processes, especially those that can have a considerable impact on people’s human rights (e.g., critical safety decisions in autonomous vehicles; allocation of health and social service resources, etc.). Further, such an integration of algorithmic systems has the potential of significant privacy and ethical consequences for individuals, organisations and societies as a whole. From a utilitarian perspective, it may be noted that transparency and accountability are both tools to

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

¹ Human Centred Systems, University of Brighton, Brighton, UK

² AI & Society: Journal of Knowledge, Culture and Communication, Springer, London, UK

promote fair algorithmic decisions by providing the foundations for obtaining recourse to meaningful explanation, correction, or ways to ascertain faults that could bring about compensatory processes. We should, however, be mindful that this notion of transparency and accountability as tools of governance may shift the accountability discourse to technical properties of algorithmic systems, for example to the inscrutability of ‘black box’ algorithms, especially in relation to machine learning systems. This shift may further increase opaqueness, thereby limiting the process of accountability. Moreover, it is worth noting that even in the realm of ‘black box’ algorithms, a key idea of algorithmic transparency is to understand not just how the system works but how it behaves also. The Study (ibid.) notes that understanding only how a system works is likely to be of little value for the transparency of individual outcomes. In that case, approaches providing explanation become more important. To pursue the idea of meaningful transparency into the behaviour of algorithmic systems, it is proposed that mechanisms for behavioural transparency may need to be designed into such systems. It is, however, recognised that while there are efforts at integration of ethics in algorithmic decision-making, there is currently no convincing evidence that these will be sufficient to provide the necessary accountability safeguards for citizens. Rather than focusing on technological innovation in transparency and accountability of AI agency, the STOA Study argues for the algorithmic impact assessment requirements for systems that have a significant impact on the public. Such requirements, it says, are in keeping with the democratic responsibilities of public service provision. At the level of direct intervention through legislative means or regulatory bodies, there is a potential for an approach that combines risk assessment by a regulatory body with corresponding levels of liability. The study highlights the need for global coordination for the assessment of governance frameworks and the development of algorithmic decision-making processes. To implement these assessments, four mutually reinforcing policy options are proposed, addressing: (1) awareness raising: education, watchdogs and whistle-blowers; (2) accountability in public sector use of algorithmic decision-making; (3) regulatory oversight and legal liability of the private sector; and (4) the global dimension of algorithmic governance. As a contribution to the understanding of AI agency moving more and more to the centre of AI discussions in areas such as automated decision-making, mediation, surveillance and facial recognition, the AI and the Future of Humanity Exhibition (2020), asks the question: “What constitutes intelligence, and if and how can intelligence be constructed by algorithms and machines?” In pondering the ‘Question of Intelligence’, the exhibition curators recognise the implication of our environment being ‘increasingly governed by machine learning processes that are looking for patterns everywhere, these

works raise philosophical questions about sentience and how our understanding of the world is changing’. In seeking investigation into social and cultural transformations generated by AI, the exhibition raises the question of what the encoding of ‘intelligence’ means for the state of being human in an environment of human–machine collaboration.

In another response of the trappings of AI Agency, Crawford and Whittaker (2018) point out that automated decision-making is already reshaping the working of the criminal justice systems work via risk assessment algorithms and predictive policing, changing our work-place culture, impacting employment, and redefining educational systems through automated evaluation tools and matching algorithms. There is, thus, a need for public agencies to urgently cultivate a practical framework to assess automated decision systems and to ensure public accountability. This framework needs to find a way to address the accompanying risks to fairness, justice and due process, and to include affected communities in the conversation. This conversation should also include issues of algorithmic accountability and liability, data privacy and security, data governance and consent to use, and transparency, bias and fairness of automated decisions as well as the impact of algorithms on human behaviour. In seeking an insight into issues of accountability, de Laat (2018) explores issues of opaqueness, transparency and responsibility of algorithmic systems and asks ‘How to install a culture of accountability?’ In seeking an understanding of these issues, the author wonders whether we should follow the ‘dominant trajectory of developing algorithms that become ever more accurate but ever less intelligible?’ or we should follow the trajectory of seeking explanations for algorithmic decisions? The author concludes that by opting for the trajectory of explanation, we may expect ‘full accountability’ but with the limited amount of transparency in algorithmic decision-making, ‘given the state of the art in machine learning and the current political climate’. It is noted that a most vexing question of AI Agency remains that ‘almost all aspects of algorithmic decision-making remain opaque.’ Not only private institutions use this opacity as a norm for ‘considering their algorithms to be their intellectual property’, but also public institutions ‘shield their algorithms from the lime-light.’ The opacity of algorithmic systems is also used by ‘banks and insurance companies, tax agencies, police, and airport security departments’ to keep the secrecy of their algorithmic decision-making under close ‘wraps’. To get a further glimpse of the hold of opaqueness of algorithmic systems, we observe that the AI-powered machine of data analytics is ‘touted as tools of digital transformation and predictive catalysts for reinventing institutions, organisations, exemplified by technological enterprises such as Amazon, Facebook, Airbnb, and Ant Financial.’

In the realm of the technocratic world, there pervades a vision of human–robot co-existence society. Dignum (2018) points out that whilst AI researchers are becoming more and more stimulated by autonomous agents and team-mates, concerns are increasingly being raised about ethical and legal implications of these systems, as they impact decision-making, raising further issues of accountability, transparency and regulation. These issues should be seen ‘more than the ticking of some ethical boxes’, and should be seen as part of wider societal concerns of civil rights, reliability and trust of AI systems. These societal concerns exhibit the crisis of AI Agency when autonomous agents and machine learning systems are integrated in human systems. In this perspective, AI Agency is seen not about designing algorithms, but how they are going to be used now and in future. A further concern is that in what ways AI Agency is going to have impact on people and society—for example not just how to design predictive policing algorithms to control a riot, but also ‘raising questions of what it would be like living in a police state, the type of society.’ The crisis of AI Agency also lies in the implication of the algorithms over-riding human decisions, replacing human judges in making judgment of who committed the crime and who gets bail. Questions arise as to whether we should trust human judges over algorithms even if humans make errors, may suffer from prejudice, and arrive at different judgments depending upon the context. But what price are we willing to pay for the algorithmic consistency, when we cannot be sure of the consistency of weights of prejudice that algorithmic designers attach to human behaviours, trust them without knowing of bugs not only in the data but in the rules and weights that govern them? Further how can we over-rule algorithms if we get mesmerised by them? Once we dress up algorithms in logical rules, quantitative weights, machine learning, it gets an air of authority, it is difficult to argue with. The danger is that if we get used to blind faith in the machine, giving up power to the machine, become dependent upon algorithms, and then we wonder whether there would be space for human intervention in societal decision-making. Before we are mesmerised by the consistency of the algorithmic systems, we should remember that human judgment demands ‘data of life rather than just sensitivity to the data set of facts, sensitivity alone does not make perfect algorithms but a combination of human specificity and algorithmic sensitivity.’ For an example, a trained radiologist would never miss a normal cell, whilst an algorithm never misses a tumour. It is the collaboration of human judgment and algorithmic calculation that makes for diagnosis. When we enter into the area of human rights and ethics, beyond the technocracy of ticking of ‘ethical boxes’, we encounter not only the issue of the interpretation of universal human rights within diverse social and

cultural contexts but also how tech companies shroud their ethical responsibility in the institution of ethics boards.

In making a contribution to the ongoing debate on accountability frameworks, Selinger (2019) alerts us to empty slogans of transparency and compliance and “ethics washing” facades of external AI ethics boards of tech companies. He notes that whilst these companies may tout transparency, they promote “trade language that prevents it”. Selinger further notes that design decisions, implementation strategies of AI systems and resulting unintended consequences have the potential to impact lives across the globe. In military, law enforcement, banking, criminal justice, employment, and even product delivery contexts, algorithmic systems can threaten human rights by automating discrimination, chilling public speech and assembly, and limiting access to information. In raising these societal implications, he argues for a rights-based blueprint for business, and data and AI governance that upholds human rights and dignity. In seeking a framework for ethics-based human rights in the age of AI, he quotes that: “Whether our ethical practices are Western (Aristotelian, Kantian), Eastern (Shinto, Confucian), African (Ubuntu), or from a different tradition, by creating autonomous and intelligent systems that explicitly honour inalienable human rights and the beneficial values of their users, we can prioritize the increase of human well-being as our metric for progress in the algorithmic age.” As a contribution to this debate on prioritisation, Nathan Strout (2019) brings to our attention the issue of investment deficit in AI oversight. He notes that although we meet rhetoric of public policies and academic debates on privacy, civil liberty, transparency, trust and ethics, the focus of AI systems and tools remains on performance in the dominant techno-centric design tradition rather than on holistic design ethos enshrined in the human-centred tradition of symbiosis. Scott Shackford (2019) alerts us that in this age of disruptive technologies of machine learning and data analytics, we find our identities being linked to facial recognition, where ‘your face is not just a unique part of your body anymore; it is the biometric data that can be copied an infinite number of times and stored forever.’ We are told that now that facial recognition algorithms exist; they can be effectively linked to any digital camera and any database of labelled faces to surveil any given population of people. For example, we are warned about the addition of a ‘robot dog’ to bomb squads, raising concerns ‘that Skynet and the dystopia of *Black Mirror* are upon us.’ This raises a deep concern of almost limitless implications of surveillance and potentially weaponization operations that may be allowed, especially when a robotic system like the robot dog is used in concert with technologies of Big Data and Machine Learning. In raising issues of ethics and trust in the use of artificial intelligence tools in making combat decisions, Freedberg (2019) alerts us to the need for cultivating

room for human judgment, creativity, and ethics in designing AI-driven systems. In the quest for a balance between disruptive technologies of data analytics, artificial intelligence tools, autonomy and the operational trust, it should be recognised that although we can leverage AI-driven systems, we cannot trust them blindly. This, however, requires that machine learning tools should be so designed as to accommodate a built-in ‘confidence factor’, with a reliability at least ‘as good as a human’, especially when we deal with data that are unreliable, incomplete, or even deliberately falsified. It is emphasised that “The art piece is that human, at a minimum in the loop, to be able to look at that data and results of [the] decision-making tool or algorithm, [and asking], ‘does this make sense?’” So the challenge is how not to lose ‘the creativity of the human mind’ and ‘how do you actually input that human factor into an AI system?’

To counter the images of a dystopic future in media and popular fiction, The Royal Society’s discussion on AI, Society and Social Good (2018) offers an argument of a positive impact and potential of AI systems for societal benefit in domains such as human health, transportation, service robots, health-care, education, public safety, security and entertainment. Whilst posing new societal questions of ethics, trust, reliability and transparency, the discussion also encourages new conversations about the ways in which AI technologies are becoming transformative catalysts in shaping the world. However, to cultivate trust and reliability, AI systems and tools need to be aligned to social, cultural, legal and moral values of those impacted by them, and this cultivation needs to be guided by ethical frameworks. It is also noted that real-world data are messy: they contain missing entries, they can be skewed or subject to sampling errors, and they are often collected for purposes other than the analysis at hand. In addition, the models created by a machine learning system can also generate issues of fairness or bias, even if trained on accurate data. In recruitment, for example, systems that make predictions about the outcomes of job offers or training can be influenced by biases arising from social structures that are embedded in data at the point of collection. The resulting models can then reinforce these social biases, unless corrective actions are taken.

In this volume, our authors reflect on some of these issues and provide an insight into the problematics and potentials of the AI-powered machine. Manuel Carabantes (this volume) notes that although artificial intelligence models with machine learning may exhibit the best predictive accuracy, they are paradoxically also the most opaque black box architectures. Commenting on the argument that ‘transparency is a design criterion that we demand from AI but not from other human beings’, Carabantes says that what it means is that we do not have access to other human beings’ minds or brains, and we do not demand it to trust them. In other words, we only have access to their behaviour and their explanations,

which are not a reliable depiction of the real mental processes. He quotes that ‘although it is impossible to access the thoughts of other human beings, evolution has endowed us with mirror neurons and theory of mind to infer how they do it; tools that have proven to be effective over millennia’. Instead, the computer is a mystery to us; a mystery that gets darker the farther it moves away from being a simple pocket calculator and the closer it gets to AI, be it symbolic or sub-symbolic. He further asserts that if AI is symbolic, then it is hardly comprehensible by the user, because its heuristic rules, which act as our cognitive biases, are different and also tend to the minimum to explore the whole space of computationally possible solutions. Inspired by the argument that artificial super intelligences (ASIs) pose an existential threat (called a technological arrest), Bradley (this volume) seeks an alternative management framework to mitigate vulnerabilities arising from misbehaving AIs. The argument is that if AI behaves in an unexpected way, there may be no safe way to bring it back online with any degree of certainty that it is not going to continue to develop capabilities and probe for vulnerabilities. Like malicious threat actors such as professional criminals and terrorists, Artificial Superintelligence will be capable of rendering mitigation ineffective by working against counter measures or attacking in ways not anticipated by the risk management process. Risk models for AI, thus, need to go beyond the belief that as humans we can somehow outsmart super machine intelligence, thereby need to shift from static, anthropomorphic models and focus on data-driven models to measure intent, manage intent and prevent the treacherous turn.

Naude and Dimitri (this volume) reflect on the existential threat of an unfriendly AGI to humanity, arising from the arms race for an artificial general intelligence (AGI), and propose an all-pay contest public procurement model for public policy to avoid such an outcome. The public procurement, it is posited, would reduce the pay-off of contestants, raise the amount of R&D needed to compete, and coordinate and incentivize co-operation. What is needed for public procurement of AI innovation is to appropriately adjust the legal frameworks underpinning high-tech innovation, in particular dealing with patenting by AI. Boyles and Joaquin (this volume) take up Muehlhauser and Bostrom’s debate on ‘Why Friendly AIs won’t be *that* Friendly’ and conclude pessimistically that we need to brace ourselves for the impending rise of intelligent machines and hope for the best, since these AIs will not be *that* friendly. This pessimism stems from the possible disconnect between ideal human values and morals and those perceived of the machine. The authors examine the embedding of our most cherished values and beliefs to the operational values of the machine. The problem with this bottom-up embodiment of values, authors argue, is that the embodied AI option do not have pre-programmed set of behaviours; rather, they are equipped

with basic behaviours that are instantiated as they interact with their environment. On the other hand, Artificial Life applies evolutionary techniques to come up with virtual, simulated intelligent life forms. Both these design strategies mimic how the intelligence level of humans was, or is continuously being developed. Moreover, the problem with embodiment is that it holds that human values would not only remain stable but also homogeneous, and that humans are by nature friendly. This to authors seems implausible given the brutality of human history. The authors say that even if we grant that our moral sphere is pretty much stable nowadays, are we ready to tread Milton's long, hard road once more to achieve AI friendliness? Towards this road to friendliness of AI, Scott Robbins (this volume) sheds light into how Luciano Floridi's concept of 'envelopment' may help us to understand the puzzle of creating AI-driven machines that will benefit and not harm society. By considering the question of what makes for successful robotics in this age of techno-optimism, the author posits that this envelopment of robotics is regarded as simply one part of the puzzle in creating AI-driven machines. This puzzle also faces the challenge of ethical evaluations for delegating tasks to a machine, while striving for a meaningful human control of the machine operation. In response to concerns about stifling innovation, it is pointed out that envelopment allows for opaque algorithms to do what they do best, thereby keeps the opacity constrained to how the machine makes decisions. However, Robbins argues that opacity which spreads well beyond the 'how' of the machine into the what, where, why, etc., is unacceptable. Moreover, to move beyond just constraining opacity would involve responsible design, development, and implementation of AI-powered machines. The argument is that the debate on opacity is very much concerned with ethical issues that arise from the 'black box' operations of modern machine learning algorithms, thus constraining the understanding of the reasoning behind algorithmic decisions, especially 'morally salient decisions'. It is proposed that rather than constraining AI-powered machines, we should allow these machines to realise their function whilst preventing harm to humans. However, it is argued that to put an 'envelope' around AI-powered machines, we need to have operational knowledge such as the training data, inputs, functions, outputs, and boundaries, and only with this knowledge we can responsibly regulate, use and live in a world populated by these machines. González (this volume) provides a Cartesian window to get a perspective of the operational limit of the AI machine. The argument is that by enthusiastically embracing part of the Cartesian view, some AI researchers have simply overlooked the mind as a phenomenon with a first-person viewpoint. This viewpoint is neither reducible to algorithms, nor to any sort of machine-like behaviour. The author argues that paradoxically Descartes' view about

reason and intelligence has encouraged certain classical AI researchers to suppose that linguistic understanding suffices for machine intelligence. However, Cartesian metaphysics involves a second part, which is directly related to whether machines can think, on the one hand, and to whether reason can be mechanised on the other hand. It is argued that Cartesian reason is flexible and unlimited, and that rational beings can in turn produce unlimited arrangements of words. In contrast, programmed machines, qua physical things, cannot produce unlimited flexibly relevant linguistic outputs. Continuing this perspective of Cartesian reason, we get an insight into the algorithmic automatisisation of experiential world. Chiodo (this volume) argues that we may be developing a novel human condition in the sense that algorithmic abstraction is always better than mental abstraction. This shift in abstraction, in the Western Culture, lies in progressive dominance of the realm of rationality, leading to the realm of computation, and then to the realm of algorithmic automatisisation. In this shift to algorithmic rationality, we are progressively externalising not only human contents but also human abilities, i.e., we are progressively atrophying ourselves by becoming creatures who are progressively delegating the core of our very essence of being. This essence has always included the epistemological ability, together with the ethical courage of making complex decisions on both our lives and the others' lives. Chiodo alerts us to a paradoxical possible scenario, where a total externalisation will make the algorithmic machines human, taking with them all the most essential human prerogatives. In turn, human beings become mechanical, thereby losing all the most essential human prerogatives.

In pursuing the discussion on what it means to be human being, we wonder whether the awarding of legal personhood and citizenship to artificial intelligence agents follows a path of externalisation of essential human prerogatives to the artificial machine. Tyler L. Jaynes (this volume) reminds us that neither conception of artificial intelligence nor the impact of granting legal protection or citizenship to AI agents is new. What is new is the notion that artificial intelligence can possess citizenship—a concept reserved only for humans, as it presupposes the idea of possessing civil duties and protections. The problem is that without developing a set of legal and moral rules regarding the treatment of artificial general intelligence (AGI) and other non-biologic intelligences (NBIs), humans cannot fulfil our moral duty to preserve the foundations of the social contract. As technology has progressed far beyond what the traditional law had intended to cover, current attempts to regulate AI technologies are making only marginal changes. The author argues that it will not be enough to state that AGI and similar NBI systems have rights, nor to enforce those rights with legal action, what we need is to consider how AGI and similar NBI systems function, and how their freedom impacts the corporations

developing these systems, before any progress can be made to grant legal protections for these intelligent entities. Beyond the notions of legal personhood and citizenship, the author argues that there is a need to develop an ethics committee akin to the partner of the Human Genome Project to consider ethical qualms and other socio-economical or political issues that may arise as policy on legal rights of artificial entities begins to be drafted. Continuing the debate on legal rights of artificial entities, Spatola and Urbanska (this volume) comment that our perception and acceptance of robots and artificial entities depend upon the representations we develop for them. These representations, usually guided by information from fiction, impact our expectations of these entities as positive or negative. As AI and robots become more and more present in our everyday life, this may lead to a more positive adaptation between human and artificial entities. However, it should be noted that science has mainly taken an anthropocentric perspective (i.e., how distant from humans are these agents), thereby influencing the way humans perceive these artificial entities. This perception can create both positive expectations and fear from robots and artificial entities, at times alluring humans to their conceptualisations as divine entities (e.g., gods). In an experiment on the perception of these entities, it was found by authors that participants' representation of artificial intelligence, robots, and divine entities was similar, while the representation of humans tended to be associated with that of animals. It illustrates that people make sense of the new entities by relying on already familiar entities and in the case of artificial intelligence and robots, people appear to draw parallels to divine entities.

In making some sense of the concerns of the AI-powered machine, Amershi (this volume) wonders whether the emergence of the AI machine can be understood within a cultural perspective that drives this technological process. And whether a change of “cultural perspective” provides us with new insights to steer the process and change this technological course. The author notes that various cultures have produced different ways of ‘thinking’ and dealing with the world. In the West, the Age of Enlightenment induced revolutionary changes in the way of thinking and liberated the Western mind from the limiting fetters of religion, a priori beliefs, preconceived truths and the related shackles of power. We can now argue that this process of ‘liberation’ in the current phase appears to have reached a point where continuation on similar lines of thinking poses fundamental threats to human existence. At this juncture what is needed is a fresh perspective on which the trajectory of the technological process should take place. The author reflects whether alignment of Western and Eastern cultural parameters, seeking an interdependence and universal harmony (between man and between man and nature), could be a viable alternative to reframe our relationship to technology

and the technological process, so that the faculties of reason and rationality can be grounded back on ethical and moral standards for assessing all our actions. Nawar (this volume) through his interactive art project on Collective Bread Diaries explores the complex relationship between technology and cultural, historical, and political realities that have long existed across the history of mankind. Drawing on the historical imprint of bread on revolutions throughout history, and common meanings it signifies to people from all over the world, the author reflects upon the meaning of machine revolution. It is noted that data collected on bread from across diverse cultures interactively in the form of Collective Bread Diaries can be seen as a collective learning machine. It is conjectured that the learning machine of the future will learn as more and more data are gathered from different sensors, enabling the capturing and controlling of complex processes. It is suggested that as the machines adapt to human needs and habits, taking over tasks of sensory organs, becoming more like humans, thereby making more complex systems accessible to humans. Złotowski et al. (this volume) explore cultural perspectives of social robots within the UAE context, and wonder whether globalisation leads to a shared perspective or universal cultural parameters that affect expectations and social acceptance of these robots. The authors suggest that the existing separation of males and females present in many public places in the UAE was exhibited in preference for the same-gender or at least the lack of preference for opposite-gender androids, and this seems to be in stark contrast with both the Far East and West. Although cultural differences may be interesting, the authors argue that it is not enough to design robots merely to optimise their human-likeness for their task performance. Factors affecting social acceptance, such as likability and non-threatening appearance, are more crucial in determining whether a robot will be preferred for a given job, which emphasises the importance of social acceptance in robot design.

Drawing from the theoretical discussion within epistemology about the differences between knowledge-how (LH) and knowledge-that (LT), Loh and Misselhorn (this volume) explore the impact of augmented learning on the acquisition of knowledge, providing an insight into the shifts in the learning focus from propositional knowledge to epistemic competencies, which can be differentiated as ‘grasping’, ‘wielding’, and ‘transferring’. These learning shifts are identified in terms of epistemic competencies, autonomous, self-directed learning, and motivation of learner’s involvements in a variety of learning environments. We learn how augmented learning can facilitate and support the already ongoing and sought-after shift from mere KT to KH, for example, how it enhances the skillful performance and competency to transfer the acquired knowledge to other contexts and apply it to problems outside of academia. Commenting on their

experiments on Augmented Learning, Smart Glasses and Knowing How (ALSG), the authors acknowledge that “it is yet unclear whether this merely adds a form of ‘gamification’ in the sense of an extrinsic motivation to the experiment, or whether the visualizations are capable to instill intrinsic curiosity and wonder in the learner.” Kraus (this volume) in the dream of the chemist asks us to reflect on the bearing of logical intelligence and critical intelligence. It is argued that AI’s logical intelligence results from the calibration of data bases of human activity, and additional intelligence can be harnessed with the strategic and logic creative insight of human intelligence, for example to discover drug molecules more quickly. On balance, the author conjectures, AI can be detrimental to human society on two grounds: first, when logical and critical aspects of human intelligence can be amplified with strong critical intelligence of which society will have no more control; and second, if human beings are lured by AI with low critical intelligence, who may misuse AI, thereby leading to a dangerous drift for human society. It may be that complex interplay of humans, society and technology would mould the threat to humanity that may arise when the amplified intelligence would reach a level of critical intelligence higher than that of human innate critical intelligence. However, Kraus alerts us to the threat from the natural stupid intelligence in the form of the emergence of artificial stupidity.

In this volume, our authors contribute to the current debates on the problematics and potentials of the AI-powered machine. The contributions raise issues of the dangerous drift for human society towards amplified intelligence, existential threat (technological arrest) of artificial super intelligences, the paradox of predictive accuracy of opaque black box architectures of machine learning, and vulnerabilities of data-driven models to measure intent, manage intent and prevent the treacherous turn. We are reminded of the opacity of algorithmic systems including concerns of ethical issues that arise from the ‘black box’ operations of modern machine learning algorithms, thus constraining the understanding of the reasoning behind algorithmic decisions. The discussions shed light on the nature of intelligence, and explore the impact of augmented learning on the acquisition of knowledge, providing an insight into the shifts in the learning focus from propositional knowledge to epistemic competencies, which can be differentiated as grasping, wielding, and transferring. Among the challenges of AI Agency, we are introduced to the challenge of ethical evaluations for delegating tasks to a machine, while striving for a meaningful human control of the machine operation. In making us aware of the implication of cultural acceptance of robots and artificial entities as divine entities, we are reminded of the implication of the shift to abstraction and algorithmic rationality that leads to the dominance of the realm

of computation and algorithmic automatisisation, thereby progressively externalising not only human contents, but also human abilities. On alignment of human values and algorithmic parameters, we wonder whether the faculties of reason and rationality can be grounded back on ethical and moral standards for assessing human–machine collaborative actions.

Once we move beyond the academic discourse on existential risk, accountability and regulatory frameworks, digital futures and rhetoric of big data, Internet of Things and deep learning, we meet Spiegelhalter (2020) who poses a central question: “can we trust algorithms?” And in our case, we may ask: Can we trust AI Agency? In asking about trustworthiness, Spiegelhalter reminds us of the philosopher Onora O’Neill (2013) who says that ‘organizations should not try to be trusted; rather they should aim to demonstrate *trustworthiness*, which requires honesty, competence, and reliability.’ Although there is an argument for claiming trustworthiness of a system in terms of its features of explainability, accuracy, auditability and responsibility, we cannot take it for granted that AI Agency or its algorithms would be beneficial to people and society, when implemented. Beyond the rhetoric of collective intelligence and digital democracy, we should expect that the design, evaluation and implementation processes of AI agency adhere to the ethos and practice of trustworthiness. Spiegelhalter further notes that we should be mindful in presupposing that ‘algorithms shown to have reasonable accuracy in the laboratory must help in practice, and explicit evidence for this would improve the trustworthiness of the claims made about the system’. Even in the case health system, he notes that there has been remarkably little prospective validation for tasks that machines could perform to help clinicians or predict clinical outcomes that would be useful for health systems. Although there is a great ethical interest and concern about the impact of AI Agency in health-care and criminal justice systems, Spiegelhalter says that perhaps a more basic issue is whether we should believe what we hear about them and what the algorithm tells us. For those interested in the design of trustworthiness of AI Agency and algorithms, he proposes seven criteria for the evaluation of claims of trustworthiness: (1) is it any good when tried in new parts of the real world? (2) Would something simpler, and more transparent and robust, be just as good? (3) Could I explain how it works (in general) to anyone who is interested? (4) Could I explain to an individual how it reached its conclusion in their particular case? (5) Does it know when it is on shaky ground, and can it acknowledge uncertainty? (6) Do people use it appropriately, with the right level of skepticism? (7) Does it actually help in practice? In seeking trustworthiness of AI system, Spiegelhalter asks us to be mindful in that “It is important not to be mesmerised by the mystique surrounding artificial intelligence (AI).”

In its tradition of the evolving discourse and conversation on art, science, technology and society, AI&Society welcomes contributions on the trapping of AI agency and the way forward to cultivating a culture of public accountability and trustworthiness of the AI machine.

“we are being electronically monitored for a set of connections – the network of associations, people, places, calls and transactions – that make up our lives”- Eyal Weizman (who was barred from travelling to the US after being flagged by the algorithm as a security risk) - The Guardian 20 February 2020) (<https://www.theguardian.com/artanddesign/2020/feb/20/head-of-turner-prize-nominated-forensic-architecture-barred-from-visit>)

References

- Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability, Dillon Reisman, Jason Schultz, April 2018. <https://ainowinstitute.org/aiareport2018.pdf>. Accessed 21 Feb 2020
- de Laat PB (2018) Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philos Technol Springer* 31:525–541. <https://doi.org/10.1007/s13347-017-0293-z> (accessed 21.2.2020)
- Dignum V (2018) Ethics in artificial intelligence: introduction to the special issue. *Ethics Inf Technol* 20:1–3. <https://doi.org/10.1007/s10676-018-9450-z> (accessed 21.2.2020)
- Freedberg SJ Jr (2019) The art of command, the science of AI. *Netw Cyber*, November 25, 2019. <https://breakingdefense.com/2019/11/the-art-of-command-the-science-of-ai/>. Accessed 21 Feb 2020
- O'Neill O (2013) What we don't understand about trust. TEDxHouseOfParliament. https://www.ted.com/talks/onora_o_neill_what_we_don_t_understand_about_trust/transcript?language=en. Accessed 27 Feb 2020
- Selinger E (2019) Human rights and AI ethics. <https://onezero.medium.com/a-i-ethics-boards-should-be-based-on-human-rights-2b4323b394e4>. Accessed 21 Feb 2020
- Shackford S (2019) Massachusetts police test out robot dogs. Is dystopia on its way? <https://reason.com/2019/11/26/massachusetts-police-test-out-robot-dogs-is-dystopia-on-its-way/>. Accessed 21 Feb 2020
- Spiegelhalter D (2020) Should we trust algorithms? *Harvard Data Sci Rev*. <https://doi.org/10.1162/99608f92.cb91a35a> (accessed 27.2.2020)
- STOA (2019) A governance framework for algorithmic accountability and transparency STUDY Panel for the Future of Science and Technology. EPRS|European Parliamentary Research Service Scientific Foresight Unit (STOA) PE 624.262—April 2019 [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU\(2019\)624262_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf). Accessed 4 Feb 2020
- Strout N (2019) Where is the investment in AI oversight, asks inspector general? *Artificial Intelligence*. C4ISRNET. June 10, 2019. <https://www.c4isrnet.com/artificial-intelligence/2019/06/10/where-is-the-investment-in-ai-oversight-asks-inspector-general/>. Accessed 21 Feb 2020
- The AI and the Future of Humanity Exhibition (2020) The Question of Intelligence—The New School, 66 West 12th Street, New York, NY 10011. <https://blogs.newschool.edu/news/2020/02/the-question-of-intelligence-ai-and-the-future-of-humanity-exhibition-at-the-new-school-explores-works-focused-on-artificial-intelligence-and-what-it-means-for-the-state-of-being-human/>. Accessed 20 Feb 2020
- The Royal Society (2018) The Discussion on AI, society and social good. <https://royalsociety.org/-/media/policy/Publications/2018/ai-and-society-workshop-notes.pdf?la=en-GB&hash=0695C536BACBF6985F01DEEE6396608E>. Accessed 2 Jan 2020
- World Economic Forum (2019). AI Governance: A Holistic Approach to Implement Ethics into AI. White Paper. 91-93 route de la Capite CH-1223 Cologny/Geneva Switzerland. <https://www.weforum.org>. Accessed 4 Feb 2020

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