AI & SOCIETY (2020) 35:783–793
https://doi.org/10.1007/s00146-020-01079-8

PREFACE

Ethics of engagement

Karamjit S. Gill[1]

Published online: 10 October 2020

In this volume, *AI&Society* authors critically reflect on ethics of engagement. The narratives range from societal sustainability, Surveillance Capitalism, Machine theology, Social jurisdiction, Covid-19, EU GDPR consent mechanisms, Strategic Health Initiative, Watson for Oncology, Recommender Systems, and Socio-technological systems. The discussion and arguments range from Artificial wisdom; Artificial moral agents; Crisis of moral passivity; Smart phones on wheels; Disengagement and re-engagement with roboethics; roboaesthetics; interpersonal interaction and perceived legitimacy; Value conflicts, Nudging traps and algorithmic bias**,** Digital Fake News, Social anxiety, and Dysfunctional impacts of automation on social and political stability; Regulatory frameworks and EU GDPR consent mechanisms; Legal, political, and bureaucratic decision-making; Implication of autonomous decision making on judgment making during COVID-19 pandemic; AI, medicine and ethics; Global supply chain dependency and Global concordance; Narrative of entanglement; AI and shared human motivations; Cognitive-architecture for autonomy, intentionality and emotion as prerequisites for creativity; Turing's vision and cooperative challenge of language use; and Theistic AI narratives.

Patrick Gamez et al. in 'Artificial Virtue: The Machine Question and Perceptions of Moral Character in Artificial Moral Agents' (this volume), investigate the "machine question" of whether virtue or vice can be attributed to artificial intelligence; that is, whether people are willing to judge machines as possessing moral character. Self-driving cars opens up the concrete possibility of encountering familiar moral dilemmas in the real world, for example, whether to save a group of children who have suddenly darted into the road or swerving to avoid that collision and instead colliding with a single pedestrian properly using a crosswalk. To authors, it is obvious a *moral* question; there is no morally neutral decision procedure here. Virtue ethics seems to be a promising moral theory for understanding and interpreting the development and behaviour of artificial moral agents. The authors explore virtuous ethics through the lens of three types of artificial agents implicit ethical agents, explicit ethical agents, and full ethical agents where *implicit moral agents* that are constrained by ethical norms even if they are not explicitly represented by ethical language; *explicit moral agents* that are capable of explicit reasoning, might explicitly represent moral rules to themselves, and use these moral rules to guide their behaviour "on the go", so to speak; and to be a *full moral agent* is to be both a moral agent *and* patient., For the authors, virtue ethics speaks of core features: rather than making *actions* the central focus of moral evaluation (as with deontology) or including *states of affairs* (as with consequentialism), the virtue ethicist takes *character* to be the primary subject of evaluation. Character, in this sense, means the set of stable dispositions, or character *traits*, to act in determinate ways responsive to features of one's environment. Using the example of "social robots" that are used to perform relational functions such that of providing empathy and intimacy or even encouragement and advice, the authors capture their view of virtuous ethics. From this perspective, moral machines and algorithms must be something like the virtuous person, or at least the person aiming to become virtuous in the sense of employing of ethical reasoning to produce ethical outcomes. The authors point out that if what matters ultimately is the flourishing of the virtuous agent, then perhaps we do not care so much about the wellbeing of the robots in question, but only about the benefits their virtues yield for us—e.g., trustworthiness, safety, etc. If so, then the virtues in question are only instrumental. They argue that even in this case, we *encounter* them in deeply social ways and wonder about their *social characters*, what kinds of characters they are, and what it would be like to encounter them. For proponents of the "social-relational" approach to the machine question, it is these encounters that matter. *If* our encounters, relations, or perceptions of these AMAs seem to be robustly, experientially ethical, then it seems that we owe them certain

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

[1] University of Brighton, Brighton, UK

treatment. They note that although humans still receive more moral blame than AMAs do in these tandems, the people are given a lower amount of blame than when they commit these moral violations on their own. As such, this could lead to the use of AMAs as scapegoats in future human moral violations and must be considered when integrating autonomous agents in the future. Although humans are attributed a higher amount of moral blame than AMAs, the fact that the machines are treated as subjects of moral blame at all is of importance. If the level of blame attributed for moral violations varies depending on the type of moral agent, then perhaps the nature of the blame is also different. The authors note that this disparity stems from the variance in perceived embeddedness in societal structures and that higher embeddedness for humans means that obeying or disobeying commands can affect attributed blame. Conversely, since AMAs are generally not associated with societal structures, they do not see these effects to the same degree. The authors argue that when considering which policies to adopt for the future, these differences in perceived social integration between the two agents must be taken into account. There will be machines employed in a variety of different ways whose behaviour must be ethically constrained. Moreover, there are already artificial agents and algorithms to whom we delegate a good deal of decision-making; these will need to make ethical decisions. In light of contemporary deep learning techniques, one promising approach for developing these sorts of AMAs is to train them to be virtuous.

Neil McBride, in 'Developing Socially-Inspired Robotics through the Application of Human Analogy: Capabilities and Social Practice' (this volume), examines the increasing expectation that robots may participate in social care and provide some relief for the increasing shortage of human care workers, social interaction with robots becomes of increasing importance. Through a case study of the interaction of a partially-sighted social worker with a support worker, the paper explores the capabilities and technical limits of sociable robots when employed within social context of extensive complexity, and their impact on social practice and policy. It is argued that socially inspired robotics involves drawing on behavioural patterns expressed in human interactions as a basis for designing robot behaviour. This pattern requires a move away from structure and function, a social configuration of a technology across a diverse milieu of locations and situations, including the configuration of the boundary beyond which the social capabilities of the robot falls short of the social variety required. It is recognised that knowledge required by a sociable robot extends beyond both the practical knowledge of navigating physical barriers and of social engagement that many humans struggle with. However, this knowledge of navigation cannot be reduced to logical rules. Moreover, this social interaction is seen as embodied and this anthropomorphism raises issues of robotic intelligence and autonomy, and thereby the issue of manipulation of the gap between the human agency and the robot. This raises ethical questions concerning the extent to which the designers of the robot are deceptive about the robot's social potential, and the robot behaviour appearing to be more socially sensitive than it actually is, tricking us into a fantasy of reciprocation. It is posited that it is not enough for social robots to compensate for human physical frailty, they must also express social capabilities. Thus it may be more important that robots behave and interact using human social patterns than they appear human in their morphology and physical characteristics. This means that socially-inspired robots should step well beyond any design of social algorithms and concentrate on the context and location of the robot in the individual, community and societal life. The paper concludes with a recommendation that if 'robots are to contribute to human flourishing and well-being, they must offer capabilities which connect with human capabilities and enable humans to make free choices about what they do to flourish. This flourishing must be set in a social context and involve social interaction because human flourishing requires social connection and communication'.

Karolina Zawieska in 'Disengagement with ethics in robotics as a tacit form of dehumanisation' (this volume) presents a working hypothesis on roboethics that emphasise 'lived ethics'. It not only incorporates formal ethical approaches into the roboticists' work but also *being* ethical in the sense of engagement with ethical reflection. The term 'ethical' essentially means 'human'. The article refelects on whether the lack of engagement with ethics within some parts of the robotics community contributes to the emergence of a tacit dehumanisation process in and outside of robotics. The challenge thus lies in actually engaging roboticists with ethical reflection and practice, rather than only the pursuit of roboethics agenda. In this sense, ethics is viewed here as an emergent phenomenon embedded in culture and everyday practices rather than only a specific discipline of inquiry. This focus on engagement and practice also concerns the actual persons (roboticists) rather than a mere field of knowledge (robotics). The argument is that such an approach along with the use of the overarching terms such as robotics, roboticists and ethics aims to allow for the discussion of a similarly all-embracing notion of 'dehumanisation' that applies to 'humans' and constitutes a core focus of this work. Another challenge is how to overcome the claims that ethics as a separate or only additional area of inquiry or discipline, it is for others (experts in ethics) who should reflect on ethical issues and come up with adequate solutions, or 'Ethics is for private conscience, or, ethics is for philosophers or clergy, not engineers'. Moreover, ethics has sometimes been subject to not only 'engineerizing' approaches but also attempts to turn it into a product or a tool serving marketing strategies and sales objectives.

What makes the use of prevalently efficiency- and profit-oriented approaches particularly problematic in robotics is the expected high degree of integration of robotics technologies into the human everyday life and their impact on our conception and experience of what it means to be human. Also, avoiding to sufficiently engage with roboethics on both the individual and organisational level does not help achieve another fundamental principle of robotics and engineering, namely that of improving human life. Since ethics and morality are viewed here as the essential human characteristics, to devalue or disengage with the ethical is to devalue and disengage with the human. Such an approach may increase the already strong tendency to apply dehumanising analogies in robotics and lead to not only redefinition but also elimination of the key human features. The article reminds the reader that dehumanisation in robotics is part of a broader phenomenon taking place in our culture and society. For example, when addressing the subject of big data and related biases towards people of colour and women, it has been argued that 'dehumanization is rendered a legitimate free-market technology project' This is why this work points to a tacit form of dehumanisation taking place in robotics, an emerging paradigm that needs to be made explicit before it can be challenged. The author argues here that considering ethical and human-centred subjects as only optional in robotics research also requires strong justification and a clear assessment of the potential consequences of taking such an approach. This is also how the underlying process of tacit dehumanisation, taking place within parts of the robotics community, can be made explicit and ultimately challenged. Another way to challenge the tacit dehumanisation paradigm in robotics is by fostering the notion of 'lived ethics'. This implies understanding ethics not just as a reflection upon a given subject but also a particular way of being in the world, a lived ethics that 'points to the mutual shaping of ideas and real life and suggests that moral systems should not simply be applied to concrete situations but rather applicable to and livable in them'. Also, this emphasises on 'lived experiences' is a key to help the integration of ethics into the actual roboticists' thinking and conduct, thereby connecting roboticists to the ethical concerns of wider society through emphasises on a shared notion of humanness. Further, by associating ethics with culture, the entire narrative around robotics technologies helps us move away from discussing ethics in terms of 'traps' or concerns towards the terms with a more positive connotation, i.e. that of 'values', and hence increase the overall individual and collective engagement with roboethics. The article asserts that roboethics should imply a long-term ethical reflection and practice undertaken within the robotics community that would ultimately lead to the change of the entire engineering culture from a predominantly technical towards more inclusive and socially-oriented ethos. This assertion recognises that technology plays a major role in our culture and society, and robotics is inherently focused on different aspects of the notion of humanness, and hence ethics. Engaging with roboethics may be one of the best alternatives we have at the moment to actually choose what future we want to have and whether it will be human. It may be that an active engagement of roboticists in multi-disciplinary research would foster this socially-oriented research in roboethics.

Danila Bertasio, in The old doom of a new technology (this volume), reflects on the artistic perspective of our engagement with robots, and wonders what makes robot designers to endow them with anthropomorphic forms, even at the risk of compromising their functionality. Although the creation of a double has appears also in the world of art, the incorporating human-like features, the drive to cover their machines with a latex coating to simulate human skin, is a new departure from the artistic forms. Endowing robots with human features certainly does not intrinsically entail applying human constraints, and the added extras of the technology available might be introduced as novel elements worthy of exploration and development. For example, what does the fact that a robotic wrist that might easily be permitted to turn through 360°, as opposed to our mere 180°, entail regarding the aesthetic correspondence between natural movement and artificial movement? It is obvious that an accentuated anthropomorphism cannot but lead to a limited, non-flowering branch of robotics. In a sense, fully human-like robots would herald a new, more spectacular phase, but one that would essentially be identical to the automata tradition of past centuries. With regard to both the potential performance and the aesthetic appearance of a robot, the determination to make it a surrogate of man would actually end up limiting its potentialities. A robot with an extendable neck, for example, would prove rather more strategic, in many practical circumstances, than would a mere simulacrum of the human body, complete with its inherent limits of movement. Bertasio argues for *roboaesthetics* as a way forwards and a basis of a common project that combines robotics and aesthetics, in which machine could participate in communicating, managing and sharing with humans aesthetic values.

Daniel W. Tigard et.al in 'Socially responsive technologies: toward a co-developmental path' (this volume) presents an empirically grounded argument in favour of some technologies being designed for *social responsiveness*. The argument is that although our usual practices will likely undergo adjustments in response to innovative technologies, some systems we encounter can be designed to accommodate our natural moral responses. For AI or robots to be *socially responsive*, the argument is that we can think of the programmed parameters of responsiveness as the *social jurisdiction*. Just as we expect of fellow humans, AI or robots will likely become able to recognize the social and emotional cues of humans within the immediate vicinity, and can use this information to better meet the present users' needs. For example, a carebot deployed in retirement communities should be attentive first and foremost to the elderly person with whom it is directly interacting. Being socially responsive in terms of recognizing human reactions includes features we would expect to see in cases where humans are taking active recognition, namely appropriate *responses* to human communications. AI systems and robots can assess any damages in the present situation (human injury, misplaced groceries, etc.) and can offer potential remedies by which the users' concerns might be alleviated. The authors make us note that this conception of social responsiveness does not entail that the systems should be programmed to exhibit human emotions. Indeed, commonly encountered AI, robots, and humans alike can be socially responsive—they can help others by aiming to achieve desired social ends—without being in (or even pretending to exhibit) any emotional state. In this way, the authors assert that their account sidesteps the worries over potential deception and manipulation, and instead focuses on the potential goods to be brought about by including some AI and robotic systems within our sphere of interpersonal interaction. The article argues that where AI and robotic systems are designed with social responsiveness, we can do away with our propensity to hold AI and robots responsible. In this case, it seems that socially responsive systems are altogether unnecessary. Further it appears highly implausible that AI and robots might be made to be socially responsive while we continue to respond with our usual attitudes and practices. To best improve HCI and HRI for the future, we would do well to consider intermediary paths of development, in which we humans will continue making adjustments in our attitudes and practices so that we can effectively cohabitate public spaces with AI systems and robots. But to accommodate us, common systems too must undergo future development, including serious consideration of a degree of social responsiveness.

Giuseppe D'Acquisto, in 'On conflicts between ethical and logical principles (this volume) in artificial intelligence', puts forward a proposition whether it is possible to identify a set of rules for data use by intelligent machines so that the decision-making autonomy of machines can allow for humans' traditional informational self-determination, as enshrined in many existing legal frameworks. The debate on machines autonomy centres on the degree to which this autonomy may expose humans to the risk of unexpected adverse outcomes. Although the debate increasingly focuses on technical design of intelligent machines and identifies possible negative long-term societal effects, the discussion also makes a case for their contribution to societal benefit. It is, however, argued that the benefits from intelligent machines will be attained only if their design is aligned with a set of broadly accepted values and ethical principles. This vision of an ethically aligned design of autonomous machines includes, at the technical level, a framework of guidelines that is inspired by ethical principles aimed at an "ethical outcome". In exploring this vision, the author discusses what ethical outcomes can designers of these machines expect. And further question is: can there be a mismatch between ethical goals and machines outcomes? The discussion moves on to ethically oriented non-maleficent artificial intelligence, i.e. a machine whose outcomes are not harmful to humans, firstly from the machine reaching the *known states* that have not yet been classified by humans as harmful, and secondly, there may exist *unknown states* or even unobservable states, that humans can only infer whether these are harmful to humans. The author notes that we can then only make machines less maleficent over time, either through experience or through prediction and inference. However, the worry is that experience and prediction may have a different weight in making future machines less maleficent, since humans have a natural attitude to discount, sometimes even considerably, the impact of future harmful events and this makes the progress in non-maleficence inherently slower. There is accordingly a need for continuous update of design constraints, which may also generate a trade-off in the long term between the costs of implementing non-maleficence and the natural human preference towards a costless and unconstrained innovation. The discussion on the principle of machines beneficence raises the issue of how to foresee future consequences of the choices of goodness and cultural values the designers make, when there is both commonality and divergence of the notions of goodness and cultural values, and how to determine a priori what is *absolutely good*. To move the debate forward, the authors put forward a number of proposals. The first proposals is to resist the temptation of thinking that it is actually possible to allocate any sort of responsibility to machines for their decision-making autonomy. The argument is that it is humans' responsibility to bring or not a machine to the "point of no-return" of its autonomy, starting from which the machine has the last say on final outcomes. Human responsibility can then be enforced using the traditional tools of economic incentives and sanction-based deterrence mechanisms. The second proposal is to promote policies that imply notification of harmful machine outcomes to a trusted centralized entity, in charge of disseminating societal knowledge on how adverse events for humans materialize. The third proposal is to foster value transparency, namely the disclosure of the criteria humans and machines apply to settle disputes whenever there is a value misalignment (for instance between the visions of public vs private goodness). The fourth proposal is to promote the design of uncertain machines. It will be interesting to see, they say, how this principle is implemented in practice and to what extent technology can be used to protect individuals from technology itself, and whether personal data protection can become the fundamental building block for the design of intelligent machines that remain subject to humans' decision-making autonomy. A final proposal is algorithmic transparency.

François Kammerer, in 'Self-building technologies' (this volume) introduces the reader to a more radical self-building technologies which might become available in a distant future. By alerting us to the possibility of cognitive enhancement, the author makes us aware of the debate on whether the process of enhancing ourselves by merging with AI would lead to a loss of selfhood, or even to the destruction of our own selves as such. Although these concerns are seen to be legitimate, the author argues that cognitively enhancing ourselves with the help of AI technology could not only make us gain intelligence, well-being, power or lifespan, it could also make us become more genuine selves, by increasing the control we have on our behaviour, as well as the coherence and the transparency of our cognitive and emotional lives. By examining the potential of two 'self building' technologies, iDiversity® and iFidelity®, to improve the psychological coherence, transparency of mental life, and control of behaviour and cognitive processes, the author articulates that control, transparency and coherence as crucial features of selfhood correlate with the degree of instantiation of selfhood. In this case, the article notes that a case can me made that iDiversity® and iFidelity® can make their users (locally) more perfect selves. As such they can be regarded as self-building technologies. The author, however, recognizes the concerns of the possibility that future technological progress might provide us with radical self-building technologies, able to transform us into super-selves, as different, maybe, from "normal" selves, than "normal" selves are from diminished selves or proto-selves.

Darja Vrščaj et.al in 'Is Tomorrow's Car Appealing Today? Ethical issues and user attitudes beyond automation' (this volume), presenta a study of societally desirable Autonomous Vehicles (AV). The study focuses on user attitudes and ethics beyond the automated driving function. It examines the attitudes of young people as future prospective users, especially their willingness to accept AI technology compared to former generations. It finds that the car manufacturers envision future AVs as digital personal assistants, guided by Artificial Intelligence and Recommender Systems features, rather than utility machines. This vision is modelled on the success of the smart phone, and online website recommender systems. It is suggested that future AV owners can use personal digital assistants, smart phones as multifunctional devices with various apps, for planning, entertainment, and socializing. Car manufacturers are increasingly following this example by imagining future cars as what might be called "smart phones on wheels." However, this vision is not aligned with user values and ethical constraints suggested by AI ethicists. It is hypothesised that despite the considerable amount of negative user attitudes on privacy and responsibility, sometimes people are willing to tolerate the negative impacts of a technology for the sake of enjoying the useful and helpful side of the technology. It is further hypothesised that prospective young users might accept handing over the control of their personal data, because they will not want to miss out on the benefits of being taken around in a personalized assistant on wheels.

Karl de Fine Licht, in 'Artificial Intelligence, Transparency, and Public Decision-Making' (this volume), introduce the reader to the notion of transparency in AI decision making using public perceptions that arises from making processes and justifications transparent. The author notes that the debate has primarily focused on how transparency can secure high-quality, fair, and reliable decisions, but far less attention has been devoted to the role of transparency when it comes to how the general public come to *perceive* AI decision-making as legitimate and worthy of acceptance. The article argues that a limited form of transparency has the potential to provide sufficient ground for *perceived legitimacy* of decisions making without producing the harms full transparency would bring. It is held that transparency as a promoter of accountability can contribute to the myth of hidden politics, where the public does not believe that it actually has access to the true decision-making process. However, even though it is intuitive that transparent institutions are preferred over non-transparent ones and that they yield higher perceived legitimacy, the article notes that recent empirical research has shown that it is far from evident that increased transparency generates trust or acceptance of public policies. In some cases, the effect can even be *negative*. The problem, the author says, is that full transparency reveals the actual reality of decision-making and that real-world decision-making rarely or never lives up to these ideals. The article proposes a framework for analyzing transparency in AI decision-making in socio-technological systems. It argues that a limited type of transparency in the form of justifications for decisions, both regarding the design of AI assistants and the decisions taken by them, has the potential to ensure more legitimacy in the eyes of the public than transparency in process. It concludes that when realizing perceived legitimacy, we should, as a default, opt for having our AI assistants explain themselves rather than open up their code etc. for public scrutiny. The same is true for the decisions of decisions-makers in the process when determining the goals and relevant considerations for the assistants.

Veljko Dubljević et.al in ‘AI in the Headlines: The Portrayal of the Ethical Issues of Artificial Intelligence in the Media’ (this volume), argue for a multifaceted approach to handling the social, ethical and policy issues of AI technology. In its exploration of ethical AI is cultivating public trust and acceptance of AI technologies, the article expresses a concern about the focus of media coverage that emphasizes the disruptive potential of AI. To counter the “hype and hope” and “gloom and doom” distortion of AI debate, the authors argue for the formulation of effective policies for increasing the public benefit of novel technologies and reducing their harmful effects for individuals and society. In their analysis of current literature, the authors identify a number of factors to formulate these policies. These factors range from encouraging public involvement, avoiding undesired results, how to regulate AI, ethics in AI, research, implementing best practices, ‘human in the loop/oversight,’ recommendations for military and law enforcement, choice and responsibility, anthropomorphizing AI, and healthcare and AI. In addition to the need for a much more balanced media portrayal of AI, the article argues for the collaboration and inclusion of ethicists and AI experts in both research and public debate as well as in the formulation of regulatory framework and policy for AI technology.

Serkan Erebak in ‘The Mediator Role of Robot Anxiety on the Relationship between Social Anxiety and the Attitude toward Interaction with Robots’ (this volume), examines a possible causal chain in which the social anxiety affects the robot anxiety, which in turn affects the attitude toward interacting with robots. It asks whether the anxiety of human–human relationships affects the anxiety when they interact with robots, and what would possible effect of this relationship be on the attitude toward interaction with robots. It is noted that understanding people’s anxiety for this interaction may help to estimate their attitude towards robots. Social anxiety is defined as the individual’s self-oriented concern that how other people will rate him/her or how communicating with another person will result. Social anxiety is observed comes in two situations, contingent (communication with another person- interaction anxiousness) and incontingent (e.g., speaking in front of a community). It is posited that the contingent social anxiety may result from attributing humanistic characteristics to robots, and this may affect robot anxiety which in turn may affect the attitude toward interaction with robots. And further whether robot anxiety shows a mediating effect between interaction anxiousness and negative attitude of interacting with robots. The authors observe that robots affect the attitude of people toward interacting with the robots; because robots may stimulate the contingent social anxiety similar to that people feel around other people.

Cheng-hung Tsai in ‘Artificial Wisdom: A Philosophical Framework’ (this volume) explores why artificial wisdom (AW) matters and how artificial wisdom is possible The article identifies key motivations of building AW systems-epistemic, survival and practical. The epistemic motivation enquires what a computational system can do to replicate a variety of aspects of human excellence and what makes such a system wiser than human beings. The second motivation for building AW systems is to secure our own survival and seeks whether an “evil” superintelligence is possible, if so then how AW can ensure the survival of the human race. The third motivation for creating AW is practical and enquires whether AW can provide guidance on what to do in our real and complicated world. The author asserts that given the distinction between wisdom and intelligence, it is not difficult to see why AW is impossible in principle. AW is, at its core, intelligent. Intelligence, whether it is construed as human or artificial, is instrumentalist in character: it is constituted by means-end reasoning, the reasoning is only about the means to a given goal, rather than about the given goal. In contrast, wisdom, regardless of whether it is construed as human or artificial, should go beyond mere means-end reasoning; an agent with practical wisdom is able to deliberate well about the final goals. Because intelligence lacks what is crucial to being wise, AW, as a kind of intelligence, cannot be wise. An agent, either human or artificial, who merely knows which things are good for human well-being can only be a quasi-wise agent. In normal situations, quasi-wisdom works fine. However, in intractable situations such as value conflicts and deep disagreements, quasi-wisdom fails. A genuinely wise agent must know further *what makes* things for good well-being, and knowing what is *genuinely* or *fundamentally* good for well-being. In other words, a genuinely wise agent knows what is the *most* worthwhile value to pursue when facing value conflicts. The paper concludes that AW is possible in principle only if it adopts specificationism about practical reasoning, rather than instrumentalism about practical reasoning. Further AW is possible in practice only if it adopts variantism about well-being, rather than invariantism about well-being.

Beth Singler in ‘Blessed by the Algorithm: Theistic Conceptions of Artificial Intelligence in Online Discourse’ (this volume) explore new religious movements where theistic conceptions of AI entangle technological aspirations with religious ones. The idea of the ‘entanglements’ is drawn from Courtney Bender’s conception, and is seen here as being entangled in social life, with history, and in our academic and non-academic imaginations, and ‘spiritual forms that have thrived and been shaped by entanglements with the secular, including its powerful engagements with science and progress’. The article further explores how ‘blessed by the algorithm’ tweets (BBtA) are indicative of the impact of theistic AI narratives: modes of thinking about AI in an implicitly religious way. This thinking also represents continuities that push back against the secularisation thesis and other grand narratives of disenchantment of technological and intellectual progress. The corpus of BBtA tweets are seen to provide us with a bounded example of theistic AI narratives online, an ethnographic moment that we can find parallels for in real-world conceptions of AI. The article concludes that this entanglement of AI and religion highlights the need for agile methodologies to explore the newer spaces where discourse on AI and religion occurs. Further, the discussion on AI and religion can involve practical questions about the future of religion and the role of religion in dealing with inequalities arising from AI and automation. The AI narratives that present assumptions about the future of religion, and the future of our agency in a super-agential world, is informative to that discussion, even if the technology is not yet at that stage (or might never reach it). This article, says the author, is an addition to larger discussion of the impact of narratives on our conceptions of AI as well as to discussion on how that AI will develop. Paying attention to real apprehensions of AI is valuable, as we seem intent on proceeding with the technology.

Tupasela and Di Nucci in 'Concordance as Evidence in the Watson for Oncology Decision-Support System' (this volume) examine the practice of using concordance levels between 'tumor boards' and a machine learning decision-support system, as a form of evidence for 'Watson for Oncology'. Concordance refers to the level to which the treatment options offered by the platform agree with the treatment options that are chosen by the oncologists. The authors also address a challenge related to the epistemic authority between oncologists on tumor boards and the Watson Oncology platform by arguing that the use of concordance levels as a form of evidence of quality or trustworthiness is problematic. They recognize that machine-learning platforms can help to identify new research findings that physicians may not have time to find out for themselves among the rapidly expanding medical literature, as well as free up time for patients themselves. Moreover, these platforms can also help to speed-up the identification of treatment options, help to reduce errors, provide cost-efficiency, help provide standardized care, as well as support oncologists through uncertainty and risk. There arises, however, a question of how to evaluate the quality of data algorithms and how should machine learning decision-support systems be evaluated in general. The pursuit of validation through global concordance levels is further complicated by discussions surrounding value-flexibility in the development of machine-learning platforms, for example medical decision-making may mask fixed and covert value judgments. They note that in the medical field, for example, patient perspectives are rarely considered in developing treatment option recommendations. This suggests a reintroduction of medical paternalism to patient treatment practices. Although platforms such as Watson may provide exciting new opportunities to help oncologists make decisions about possible treatment options at a global level, there is a risk that such platforms may also introduce values and practices, which are not locally shared by physicians and patients alike.

Silvia Milano et.al in 'Recommender Systems and their Ethical Challenges' (this volume) offers a map and an analysis of the main ethical challenges posed by recommender systems, and highlights gaps in assessing their ethical impact. It concludes with the articulation of a comprehensive framework for addressing the ethical challenges posed by recommender systems. The discussion articulates ethical issues of the *design*, *deployment* and *use* of recommender systems, and the trade-offs between different interests at stake. It highlights the way recommender systems collect, curate, and act upon vast amounts of personal data, thereby shaping individual experience of digital environments and social interactions. The article notes that research into the ethical issues is still in its infancy, and is fragmented across different scientific communities as it tends to focus on specific aspects and applications of these systems in a variety of contexts. Seeing the ethical debate from a perspective of morality, the discussion identifies two classes of variables that are morally relevant, *actions* and *consequences*, and in particular *intentions*. It notes that the value of some consequences is often measured in terms of *utility* they contain. While the concept of utility can be made operational using *quantifiable* metrics, rights are usually taken to provide *qualitative* constraints on actions. The article further notes that whilst the ethical impact of recommender systems may be *immediate*, they may also expose the relevant parties to *future risks*, these include breaches of a user's privacy, anonymity breaches, behaviour manipulation and bias in the recommendations given to the user, content censorship, exposure to side effects, and unequal treatment and a lack of trust. The authors note that although currently *architectures* of *algorithmic policy* aim to mitigate privacy risks, they may constitute a mere shift in responsibility, placing an undue burden on the users. User privacy would thus not only need to take into account the likely trade-off between privacy and accuracy, but also fairness and explainability of algorithms. A possible way forward is to develop a macro-ethical approach that would consider ethical problems specifically related to data, algorithms, and practices, as well as, to how the problems relate, depend on, and impact each other. It is further noted that recommender systems can encroach on individual users' autonomy, by nudging users in a particular direction, and trapping them to certain types of contents, or by limiting the range of options to which they are exposed. The nudging traps can only be effective if their creators understand and work with the target's world view and motivations, in the sense that the autonomous agency is effectively exploited, rather than being negated. To mitigate the impact of nudging traps of recommender systems requires engaging with users, not just how users can escape *from them*, but also how users can make the traps work *for them*.

Rodolfo Leyva in 'Testing & Unpacking The Effects of Digital Fake News On Presidential Candidate Evaluations & Voter Support' (this volume), expresses a growing worldwide concern that the rampant spread of digital fake news (DFN) via new media technologies is detrimentally impacting democratic elections. The author argues that the potential electoral impact of digital fake news (DFN) in the USA, although concerning, is strongly conditional on a reciprocal interaction between message receptibility and a 'pre-existing right-wing ideological orientation'. It is noted that although the effects of news media on political opinions and behaviours in particular, are strongly dependent on several situational contexts and individual-level characteristics, the most relevant to the present article are the believability and partisanship of the news message. The article examines current frame and priming theory research, and notes that the way in which news is presented (i.e., a news report's hedonic tone and message framing), may considerably shape the ways that audiences construct and/or employ their cognitive-affective schemas of a given attitude object. This can then prime and direct people's decisions on subsequent-related judgment tasks.

Fady Alnajjar et al. in 'Can a robot invigilator prevent cheating? (this volume) present a case study on the influence of robot presence in the classroom on the morality and behaviour of students. The authors investigate whether the robot was able to deter students from cheating and maintaining their discipline, in an examination scenario, in comparison to a human invigilator or when there was no invigilator present. They observe that while explicit cheating rarely took place across all conditions, the students were significantly more talkative when they were invigilated by a robot. In conclusion, the authors discuss and speculate upon some of the ensuing implications towards not only the application of robots in education but also consequently the wider issue of the preservation of morality and ethics in a classroom in the presence of an agent.

Berman Chan, in 'The Rise of Artificial Intelligence and the Crisis of Moral Passivity' (this volume), introduces the reader to John Danaher's notion of "moral patiency", that states 'that the rise of AI and robots will dramatically suppress our moral agency and encourage the expression of moral passivity.' The article examines Danaher's overall argument strategy: First, the deprivation of moral agency otherwise found in employment. Second, after being shut out of the labour force, we might think that we can nonetheless exercise our moral agency in some other remaining arena—legal, political, or bureaucratic decision-making. But due to emerging trend of employing machine-learning algorithms for decision-making, this second arena is no longer available to humans. Third, the final and only remaining arena we might turn to is to exercise moral agency in our personal lives and relationships. However, this option of moral agency will be all but eroded, as robots as personal assistants would not only make personal decisions for us but also supply the motivation to follow-through with action by cajoling and rewarding us. The article argues for the cultivation and the strengthening the exercise of moral agency that counter the temptation of people to succumb to the temptation of over-reliance on AI personal assistants.

Neufeld and Finnestad in 'In Defense of the Turing Test' (this volume) argue that contrary to the competitive tactics, the Turing Test is the cooperative challenge of using language to build a practical working understanding, necessitating a human interrogator to monitor and direct the conversation. Since ambiguity in language is ubiquitous, open-ended conversation is not a flaw but rather the core challenge of the Turing Test. They outline a statistical notion of practical working understanding that permits a reasonable amount of ambiguity, but nevertheless requires that ambiguity be resolved sufficiently for the agents to make progress. The authors begin with an assumption that language is a proxy for intelligence, the 'best mirror of the human mind', the means of a practical 'working understanding' in a collaborative/cooperative process of intelligent engagement. This view posits two ideas, the first is idea of *practical certainty* in the sense that it's has the probability of exceeding some threshold of belief, and the second is idea of a *working understanding* in the sense of the knowing of the desired outcomes that are *sufficiently* similar to the understanding of intentions of the other, such as that of intimacy or closeness in the sense of proximity. The authors pay tribute the Turing's vision, saying that he saw in a theoretical machine that could communicate with humans in meaningful ways, even if today's talking machines remain wide of the mark. They, however, recognize that the relational and dialogic character of human knowing (and human life) has a richness and complexity in their argument on Turing Test.

Bernard Arogyaswamy in 'Big tech and Societal Sustainability: An Ethical Framework' (this volume), discusses dysfunctional impacts of automation on social and political stability, for example the way the outsize bargaining power of virtual technologies have changed both the way we think and envision our sense of self. Further, how machine learning and artificial intelligence are posing threats to individual freedoms and rights, societal cohesion and harmony, employment and economic wellbeing, and trust in democracy. Although the author recognizes that artificial intelligence (AI) based on ever-deeper neural networks has the capability to transform medical care, revolutionize transportation, enhance security using sensory recognition, and provide customized education, he argues that immediate benefits of artificial intelligence should not blind us to the extent to which individual rights, social justice, and the common good are likely to be harmed. The author further notes that regulations may do little more than slow down the damage to society and proposes the use of an ethical calculus for negotiating and setting boundaries of how AI would be used to safeguard users' and other stakeholders' interests, rather than engaging in a race to gain the most financially, militarily, and politically. It is paradoxical that technologies such as the Internet, personal computer, and smart phone, which ostensibly enable greater decentralization, have now resulted in a higher concentration of power in the hands of mammoth firms like Google, Facebook, Amazon, Microsoft, and Apple, or in the hands of authoritarian governments. With this centralization of power, we may enter a new kind of economy that has been termed surveillance capitalism or a surveillance state. The article reflects a concern that civil society and democracy, as we know, may prove to be unsustainable, and concludes that it is imperative we reflect fully on the extent to which we are vesting our technologies with power over our lives both now and into the future.

Erez Firt in 'The Missing G' (this volume), examines the set of abilities that current AI systems lack and whose implementation will result in a basic AGI system, and considers different approaches, including a hybrid one, to a comprehensive solution for an AGI. The AGI system, presented here, is an autonomous system in the sense that it can learn in an unsupervised manner, i.e. without being instructed what kind of model it should follow and what parameters or features it should extract from its surroundings. It understands the world around it in a sense that allows it to realize how to model a new problem, learn from experience in the sense of sharing and transferring the insights learned between different problem-domains, and use abductive reasoning in a way that will enable it to reach decisions and take actions based on uncertain and limited data. Much like humans, the AIG agent must be able to associate a single new observation with an already-known relation (as we do, when we observe an occurrence and classify it as an instance of the cause-effect relation) and act accordingly. To accomplish this and more, such an AGI system should be creative in at least a limited sense. It should be equipped with fundamental capabilities, namely the ability to learn anything independently, the ability to understand its domain and surroundings in such a way that enables it to extract essential features correctly to model the problem, and the ability to reason based only on uncertain and partial data, i.e. to form hypotheses and explanations, decide which of them optimally suits the situation and act accordingly. The author considers autonomy, intentionality and emotion as prerequisites for creativity, and tentatively maintain that emotions (and consciousness) are not necessary for creativity, and that intentionality is an essential property of ideas and artefacts. The article examines two approaches that integrate these capabilities into a whole system architecture. One approach is referred to as the cognitive-architecture approach. It emphasizes the tight integration of the fundamental capabilities and other cognitive mechanisms. The other approach, the brain-emulation approach, stresses the need to learn from the human brain, and it suggests that we do so by emulation. This feature is common to both approaches—they both turn to the human brain in search of insights, solutions and ideas. The author argues that both approaches may be dependent on advancements in other areas such as cognitive science, hardware development and perhaps quantum computing. The proposal is to pursue more innovative and *hybrid* systems that combine these two approaches in a manner that highlights the strengths of each approach and mutually compensates for their weaknesses.

Jean-louis Kraus, in 'Artificial Intelligence applied to the production of high added-value Dinoflagellates Toxins' (this volume), introduces the reader to the application of artificial intelligence in making faster, cheaper and more accurate DNA sequencing, thereby gaining perspective on a particular genetic blueprint that orchestrates the whole activities of a given organism. It is argued that since artificial intelligence techniques are increasingly being used as alternatives to more classical techniques to model environmental systems, AI techniques could also be applied to biological systems production. This discussion on AI is related to biological activities not only in pharmacological and medical fields, but also to its promise as a tool for chemical biology. We learn that despite the recognized value dinoflagellates as one of the rich biotechnological source of biotoxins, scarcity of such biotoxins remains a major issue, that new marine natural products start-ups have to face economically and in terms of viability. The problem, we are told, becomes even more complex since these high value biotoxins are mainly found in dinoflagellates species which are extremely fragile microalgae. To circumvent those problems, available AI technologies based on learning neural networks, could be applied at each phase of biotoxin production: chemical synthesis and hemisynthesis, biotoxin structural identification, bioreactor engineering systems, biological pathways identification through marker-passing algorithm. The authors posits that AI clearly appears as a promising tool to help new start-uppers to jump in the restricted biotoxin market, in proposing not only biotoxins at reasonable prices but also allowing the discovery of new drugs, considering that dinoflagellates marine organisms, are the sources of several thousand drugs of interest, which remain to be discovered.

Mihai Nadin in 'Aiming AI at a Moving Target' (this volume) gives an insight into the dominant view AI, based upon a skewed understanding of both AI and medicine. The author says that there is no doubt that new computation methods would help those in medical care effectively navigate the rapidly expanding acquired expertise. However, by the nature of the medical profession, physicians exchange information and experiences because the outcome, in ideal form, is life, not a competitive edge in adjudicating profit or monetizing some new ways to help patients. Medicine is focused on what is needed to maintain life. It is an endeavour within the larger context of social organization of productive activity, of economic and political interaction, and of culture. To automate activities that engage intelligence is not the same as making intelligence available. For example, in the medical practice, the physicians understand the associating symptoms with possible causes against the background of the patient's personal narration *before* they act, not necessarily act in a manner that afterwards *seems* intelligent. This understanding is context dependent and predates the action, i.e., the treatment. That is, understanding based on information is anticipatory. For this understanding to arise, intelligence is a process, there are quantified aspects (measurements) to be considered; there are also qualitative assessments to be made; and there is the empathy in the sense that a doctor experiences what the patient is going through. The pain of the others becomes the doctor's pain. They die with those dying in their hands. To ascertain that empathy will again be made possible when AI takes care of tasks that can be automated is indicative of "machine theology": we made them, they can replace us, provided that we join the "church" (or the cult, as deep learning has become). The author recognizes that of course, AI can fully automate the burdensome bureaucratic overhead of regulations and free the physician from the tasks of typing or voice inputting to recording devices. But even for this worthwhile task, the dangers of abandoning privacy, which medicine has so far protected, are real. Medicine should not begin with measuring more and more, but *with prevention*. This very simple premise can mean many things, among them, the extreme: measure everything every time. It is a sign of responsibility that there are voices warning against the consequences of creating dependencies, some of which can lead to harm. The author asserts that medicine and ethics cannot be separated: pathogenesis and ethos are co-substantial. On the other hand, the amount of dedication and enthusiasm of those who examine the new opportunities is encouraging. New ideas come to the fore; experiments are conceived and carried out; the optimism inherent in science extends into the medicine of the time of AI and of many other scientific and technological innovations.

Roberto Musa Giuliano in 'Echoes of Myth and Magic in the Language of Artificial Intelligence' (this volume) argues that to a greater extent artificial intelligence has always been entwined with the fictional. Its language echoes strongly with other forms of cultural narratives, such as fairytales, myth and religion. The author presents examples that illustrate how these analogies have guided not only readings of the AI enterprise by commentators outside the community but also inspired AI researchers themselves. The article pays a particular attention to the similarities between religious language and the way in which the potential advent of greater than human intelligence is presented contemporarily. It then moves on to the role that fiction, science fiction most of all, has historically played and is still playing in the discussion of AI by influencing researchers and the public, shifting the weights of different scenarios in our collectively perceived probability space. The article sums up by arguing that the lore surrounding AI research, ancient and modern, points to the ancestral and shared human motivations that drive researchers in their pursuit and fascinate humanity at large. In doing so, the article highlights the interrelatedness of literary fiction, myth and religion with the theorizing and dissemination of AI ideas, and explores the narrative of entanglement AI and the wider culture. This lore of the narrative of entanglement where AI meets the wider culture should serve to amplify the call to engage ourselves with the discussion of the potential destination of the technology of AI. It draws on the work of a number of scholars such as Latour and Melzer in making an argument for a fuller picture of culture to enrich our understanding of the field of AI; going sometimes beyond the well travelled path of mainstream science towards gaining insight into the field through esoteric means; seeking insights into the field of AI not just from books and papers but also from novels and films as well. The stories of the lore, ancient and modern, surrounding AI research, feed the tacit commitment of researchers that drives their quest to expand the discipline. It is this tacit commitment that renders visible the aesthetic attractiveness of the topic and thus draws attention of newer audiences. The author emphasizes that this attention in no way should be seen to invalidate the very real concerns of those who are leading the discussion on existential risk of AI. The article concludes by drawing our attention to the wisdom attributed to the baseball catcher-cum-philosopher, Yogi Berra Yogi Berra, that 'predictions are especially hard when they involve the future.'

Imine and Joyee De in 'Consent for Targeted Advertising: The Case of Facebook' (this volume), review consent mechanisms of the EU GDPR, where the consent requirements can give users a false sense of control of their data that is controlled by social media enterprises such as Google and Facebook, thereby encouraging them to allow the processing of more personal data than they would have otherwise. The article concludes that Facebook's Ad Consent Mechanism does not satisfy some of these features such as informed, freely given, clear a rmative action and explicit consent. The authors argue that although GDPR recognizes consent of the data subject as one of the legitimate grounds of data processing, consent mechanisms for targeted advertising has received very little attention from privacy researchers.

In a timely article, 'Algorithmic bias: should students pay the price?', Helen Smith (this volume) makes us aware of the entire fiasco of the use of algorithms for automatic grading of student awards in England. As a result of public pressure, the unfairness of algorithmic allocation was recognised and over-turned. There was a reckoning of the unfairness of biased algorithms as the worry was that had the biased algorithms been deemed acceptable this year, then that precedence risked their continued use in subsequent years; this would have perpetuated a year-on-year growing divide between private/independent schools and those from disadvantaged backgrounds; possibly leaving thousands of young adults behind for no good reason. A bad algorithm applied universally is awful enough, but a bad algorithm applied inconsistently is horrendously unfair to those it affects and sets a poor example to other countries and organisations looking to manage their examination system under similar conditions. The AI community is well placed to pass comment on the construction and use of algorithms; we have a responsibility to use our experience to predict, speak out, and amplify negative issues experienced by affected stakeholders.

Richard Ennals in 'A Strategic Health Initiative: Context for Coronavirus' (this volume) remind us of the Coronavirus pandemic yet another global "Kodak Moment" of disruption, and wonders why we as society shirk our responsiblity of learning from past experiences. He refelcts on his 1986 blueprint of Strategic Health Initiative", that arugued for a framework of Strategic Health, drawing on progress in medical science, advanced computing and social administration. The argument then was that advanced technologies should be developed for prevention than just for cure. For him, not much seems to have chnged since 1986. Intelligent computer technology still places a new burden on us to determine the kind of society in which we choose to live. As in 1986, Ennals still believes that researchers prefer to work on projects they believe in. Their brains cannot simply be hired for whatever purpose. Socially committed research community recognises the benefit from the motivation of work in "advanced technology with a human face. In this spirit Ennals again suggests an initiative to tap this supply of idealism. He again argues for a strategic focus for the next stage of development of an infant generation of technology, to the benefit of society in general: a Strategic Health Initiative.

Ettore Settanni in 'Those who do not move, do not notice their (supply) chains—inconvenient lessons from disruptions related to COVID-19' (this volume) sums up the inconvenient lessons of global supply dependency in the times of crisis of global dimensions such as the Covid-19. Taking the example of disruption of medical and health care supplies during the Covid-19 Pandemic, Settanni argues for a supply chain framework that focus on the choice of sourcing locations, the establishment of supply dependencies through outsourcing and offshoring decisions, and matching supply to a changing, often erratic demand.

Jeff Malpas in 'The Necessity of Judgment (this volume) discusses the implication of autonomous decision making on judgement making. He argues that what is enshrined in this conception of 'nonhuman' judgment is just the idea that judgment is itself a matter of computation or calculation as it operates over quantitative values. This conception of judgment seems to be based two assumptions, firstly that qualitative values are amenable to quantitative reduction, and secondly that an action-guiding judgment can be derived merely from an accumulation of facts, information, or data. It should, however be recognised that judgment has an inescapable *indeterminacy* about it—there is always more than one way of judging that is supported by the evidence available. Judgment is not reducible to calculation, computation, to algorithm or rule. Malpas alerts us that one of the great dangers of automated decision-making systems is precisely that they seem to present the possibility of judgment without responsibility. One cannot escape judgment—nor can one escape the responsibility that goes with judgment. The divorce of judgment from responsibility that automation thus achieves is one of its dangers, but it is certainly not the only one. Equally important is the loss of a sense of judgment as itself inescapable—judgment and the burden of judgment is at the very heart of human life. For Malpas, the desire to escape that burden is itself representative of a desire to escape from our own humanity. In our current situation, in the face of the COVID-19 pandemic, looming economic disaster, and the ever-increasing threat of climate catastrophe, a recovery of our humanity, and so of the necessity and responsibility of judgment, is perhaps more important than ever.

In its tradition of hospitality to diversity of argument and narrative, *AI&Society*, welcomes contributions on ethics of engagement that pave a way forward to cultivating a culture of human-centred perspectives of the AI machine.