AI & SOCIETY (2021) 36:1–7
https://doi.org/10.1007/s00146-020-01135-3

EDITORIAL

# Ethical encounters

**Karamjit S. Gill**[1]

Accepted: 10 December 2020 / Published online: 15 January 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd. part of Springer Nature 2021

As we encounter the challenges and dilemmas of ethics in the AI discourse, our authors remind us of the centrality of ethics and morals in reflecting on the ways AI technologies, for example that of prediction, impact the academic discourse on societal issues such as those of surveillance, facial recognition, and identity. This discourse stimulates the engagement of the journal readers in the design and evaluation of AI systems and their impact and implications for societal concerns including those of AI governance, autonomous decision-making and social robotics. For example we learn about the place of patients, social users, health care professionals and care workers in the development and evaluation of social robots, and the ways these health care actors are engaged as participants in determining the future robot use and its users. Those concerned with the implication of machine ethics inquire into the differences between the techno-centric and the philosophical perspective on ethics and how these differences influence and affect AI governance, including questions such as what we mean by the adoption of trustworthy AI systems. These concerns make us aware of the dangerous and slippery path of the shift from philosophical to legal arguments of AI governance. We wonder whether this shift in machine ethics implies that the human is no longer at the centre, but is subject to the ethical imperative, and further whether it also suggests that 'human' itself is an ethical category and not biological at all. If this is the case then the dilemma is how to assign ethical responsibility to the social robot and in what ways it can be subject to an ethical obligation, and further how it can make ethical judgments and how those judgments can be judged ethically. In the exploration of cultivating public trust and acceptance of AI technologies, there is a concern about the focus of media coverage that emphasizes the disruptive potential of AI. Seeing the AI debate from a utilitarian perspective for social good, we may see ethics as the maximization of wellbeing as identified by consequences of actions and intentions. However, this approach does not deal with justice and human rights, let alone diversity and difference. There is an argument that a moral agency could act as a constraint to utilitarianism. However, this option of moral agency may 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 nudging and rewarding us. There is also an argument for cultivating and strengthening the exercise of moral agency that counters the temptation of people to succumb to over-reliance on AI personal assistants.

In raising concerns of autonomous decision-making from a human-centered perspective, the challenge is how to keep the human-in-the-loop and shape AI systems that create and enhance symbiotic collaborations between humans and machines. This raises questions about the ethical gap between decision-making and intention, and between responsibility and intentionality. And further, whether it is governance or autonomous decision-making, we are also made aware that ethical, moral or aesthetic choices are not made by the machine, but by those who use them as ethical, moral or aesthetic agents. In exploring the issue of AI governance, we may ask whether a utilitarianism machine can accommodate various societal needs, and if so then what are the computational limitations to align AI systems to these needs? Our authors continue to engage the journal readers in the academic discourse on the rhetoric of 'legitimacy' and acceptability that ascribes understanding to machine systems. Questions also arise as to whether socio-technical systems can fill the gap between social and political opportunities and the demands of rigour of academic research. In other words, in what ways can researchers use the academic research framework to proceed from the 'descriptive level to the explanatory level and then to the level of intervention, reconciliation and reformation'?

AI&Society authors, in 'Ethics of Engagement' (Gill 2020), 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

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

[1] Professor Emeritus, University of Brighton, Brighton, UK

moral character. By using the example of "social robots" that are used to perform relational functions such as 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 ethical reasoning to produce ethical outcomes, e.g., trustworthiness, safety. Moreover, there are already artificial agents and algorithms that are delegated to decision-making, and this raises a dilemma for ethical decision-making. Socially inspired robotics further raise issues of robotic intelligence and autonomy, and thereby the issue of manipulation of the gap between the human agency and the robot. We learn about ethical questions concerning the extent to which the designers of the social robot are not explicit about the social potential of the robot in the sense that the robot behavior may appear to be more socially sensitive than it actually is, thus nudging us into a fantasy of reciprocation. 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. 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'. This perspective of lived ethics implies that we would continue making adjustments in our attitudes and practices so that we can effectively cohabitate public spaces with AI systems and robots. For the AI research community the challenge thus lies in actually engaging 'roboticists with ethical reflection and practice, rather than only the pursuit of a roboethics agenda'.

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", recognizing that there can be a mismatch between ethical goals and machines outcomes. 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*. These alignments raise the question of coherence and the transparency of our cognitive and emotional lives. Here, control, transparency and coherence are seen as crucial features of selfhood that correlate with the degree of instantiation of selfhood. However, even if full transparency were envisaged to reveal the actual reality of decision-making, the real-world decision-making rarely or never lives up to these ideals. In this case, there are concerns of the possibility that future technological progress might provide us with radical self-building technologies that are able to transform us into super-selves, and in the bargain turning "normal" selves to "diminished selves" or 'proto-selves'. 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.

In this volume of AI&Society, our authors continue to explore and examine the broader issues of ethics and responsibility. Luís Moniz Pereira, in 'The Carousel of Ethical Machinery' (this volume), discusses the impact of the cognitive revolution on social and economic problems, and that the way this revolution was brought about by the development of AI also directly impacts humanity. The author notes that it is therefore vital and urgent to examine AI from a moral point of view. The moral problems are twofold: on the one hand, those associated with the type of society we wish to promote through automation, 'complexification' and power of data processing available today; on the other, how to program decision-making machines according to moral principles acceptable to those humans who will share knowledge and action with them. The impact of cognitive revolution is that the traditional discussion "What is humanity?" is now replaced by a powerful and challenging problem around what is desirable, possible or likely for humanity to become given the anticipated crossbreeding of AI, genetic engineering and nanotechnology. From the viewpoint of action criteria, the morality perched from the sky of the past is confronted with a new perspective on the rising moral systems studied in evolutionary psychology and deepened through testable models in artificial scenarios, as is now allowed by computers. The author posits that as research proceeds, we can better understand the processes inherent to moral decisions, to the point that these can be "taught" to autonomous machines capable of manifesting ethical discernment. Thus, the problem of computational ethics becomes urgent as the knowledge ecosystem is greatly enriched by machines with increasing ethical impacts, as machines become active players in dimensions that, until now, have been attributed exclusively to humans. The author warns that if we do not exercise appropriate foresight, we can imagine the outlines of a future that will not be promising. Should the future reflect lessons learned, the myth of 'Pandora' may be replaced with an evolved version in which the irreversible cost of catastrophic error due to reckless optimism may be recognized from the start.

Attay Kremer in 'Computers Do Not Think, They Are Oriented in Thought' (this volume), argues that the computer

is not a thinking machine, nor should its competence be taken for intelligence. Rather, the computer upholds the inhuman demands of Kant's humbling resolution of Humean skepticism—"stay in your lane!" It does not need to stray and leave the domain of possible experience. It is not guided by a subjective need to judge; it is a device used to expand the domain of experience by the careful layering of mediating representations (middleware) prepared for experiential use. In order to act in this space of orientation computers require external assistance—they do not think by themselves. But the computer mediates between abstract concepts and reality and thus serves to orient thinking. In orienting thinking, machines are vastly more competent. While the intelligence of humans is a magnetic to its own compass, computers abandon intelligence for perfect command of orientation, exhibited in the refusal to follow unexecutable commands. However extraordinary computers are, and however powerful they are as intellectual and even philosophical tools, to assign intelligence to them is an anthropomorphization and an insult to what they actually do. To call machines intelligent is to rob them of their greatest asset—their stoic, oriented calculations in face of the vast conceptual horizon. Computer are too disciplined and restrained to be intelligent.

Danila Bertasio, in 'The old doom of a new technology' (this volume), asks what makes designers of anthropomorphic forms of robots, not content with constructing useful machines, to incorporate human-like features? It may that the utopia of the creation of a double is appropriated from the world of art that also coincides with the abandonment of the naturalistic imperative in various periods of exploration and innovation. Although endowing robots with human features does not intrinsically entail applying human constraints, this addition of human features may induce novel elements worthy of exploration and development. 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. From an aesthetic point of view, on the other hand, it cannot but be conceived of in terms that are wholly different from those of the ordinary man, thus creating, in this case, amongst other things, an interesting parallel with the poetic art of Amedeo Modigliani.

Daniel Varona et al., in 'Machine learning's limitations in avoiding automation of bias' (this volume), discuss the use of predictive systems in socially and politically sensitive areas such as crime prevention and justice management, crowd management and emotion analysis, and implications of misclassification, for example for the case of conviction risk assessment or the decision-making process, when designing public policies. By identifying current gaps in fairness achieved within the context of predictive systems, the authors indicate that machine learning has some intrinsic limitations which are leading to automate the bias when designing predictive algorithms. They thus argue for the exploration of other methods or the redefinition of the way current machine learning approaches are being used when building decision making/decision support systems for crucial institutions of our political systems such as the judicial system. One such example is the emerging regulatory framework that is being built internationally, based on the International Human Rights Law. This might help to define Fairness as a quality characteristic which can be further integrated into the software development process and provide software engineers with a set of good practices, methods and techniques for targeting fairer decisions from early design stages.

Stephan and Klima, in 'Artificial Intelligence and Its Natural Limits' (this volume), revisit Adler's distinction between perceptual and conceptual thought, and discuss his narrative that while humans share perceptual thought with higher animals, only human beings could perform conceptual thought. The authors note that if 'Adler was right, and something non-material is necessary for conceptual thought to take place, the prospect arises that the lack of conceptual thought will pose a limit beyond which AI systems, no matter how complex or cleverly designed, will not be able to go'. The authors further assert that AI systems will never be able to engage in the process we term conceptual thought when it takes place in the human mind. But an equally important question that remains unanswered is whether humanity will be satisfied with non-conceptual simulated thought, and lower its expectations so that the issue becomes moot.

Adam J. Andreotta, in 'The Hard Problem of AI Rights (this volume)', explores the problematics of grounding 'rights' of AIs (e.g. robots and other artefacts) in terms of superintelligence, empathy, and a capacity for consciousness. Focusing on the 'Hard Problem' of consciousness, the author argues that we cannot make the same kinds of assumptions that we do about animal consciousness, since we still do not understand *why* brain states give rise to conscious mental states in humans. However, it is argued that progress can and has been made on the problem of animal consciousness, and in turn animal rights, without the possession of a solution to the 'Hard Problem' of consciousness. By comparing animals' behaviors, and the internal mechanisms that give rise to those behaviors, with our own, we can make well-grounded assumptions about what their mental lives are like. In terms of rights, it does not matter how intelligent a creature is, or how much empathy we feel for that creature. What matters is, whether they can experience pleasure or pain. The author asserts that the creators, investors, engineers and governments who are seeking to build complex AIs, as well as society at large, should not limit their concerns to ones involving our own wellbeing. If we create conscious AIs (whether intentionally or inadvertently), we will need to take into account their interests, e.g.

by attempting to avoid inflicting unnecessary suffering on them—the same goal that is sought by proponents of animal rights. The question of what resources should be allocated to achieve such ends is a complicated one, and will depend on what kinds of AIs we create. The decision to make AIs that are conscious is not the one that should be taken lightly.

Samuel Segun, in 'From Machine Ethics to Computational Ethics' (this volume), presents a discussion on computational ethics in terms of its 'great value' and an 'important frontier' in developing ethical artificial intelligence systems (AIS), and particularly robot and machine ethics. The discussion includes debates on moral justification of creating intelligent systems, the socio-ethical and socio-economic impact that human–machine interaction may have on society, the possible conflict of human and robot rights, and the moral status of artificial intelligence systems. It is noted that attempts at programming ethics into AI systems, building an artificial moral agent, and simulating consciousness in machines fall under computational ethics. The author argues that the use of the term 'machine ethics' in literature is too broad and glosses over issues that the term computational ethics best describes. It is posited that computational ethics is a distinct, often neglected field in the ethics of AI. In conclusion, the author proposes that there are two ways to consider the import of computational ethics to the ethics of AI. One way is to embrace it as an instructive piece and encourage collaboration among ethicists, roboticists, and computer scientists. The other is to reject it, which would imply that we continue to work in silos, each to her/his own. This level of appreciation of ethics allows moral philosophers to be experts at laying the theoretic foundation upon which the computer scientists and roboticists can begin experimentation. At the same time, it affords computer scientists meaningful insight into ethics in the bid to build safe AI.

Sekiguchi and Hori, in 'Empirical studies on the effect of an ethical design support tool' (this volume), study the ethical aspects in engineering design, focusing on two functionalities: semi-automatic generation and scenario path recommendation. The authors argue that by using the scenario path recommendation, designers can update their research themes after considering the ethical impacts of those themes on stakeholders. They suggest that both functions are realized by exploiting a knowledge base of ethical and technological discourses. Further, the ethical design theory is updated based on some unexpected results of the user studies with regards to the cyclic relationship among theory, tools (i.e., experimental equipment), and observed data.

Simon Tremblay et al., in 'From Filters to Fillers' (this volume), discuss the implication of "Snapchat dysmorphia", a variant of body dysmorphic disorder, and related body image distortions that are fueled by automated selfie filters and reflect unrealistic sociocultural standard. The authors suggest that these disorders involve dysfunctional self-modelling which entails maladaptive internalization of sociocultural preferences during adolescent identity formation. Identity formation is hereby described as cycles of interpersonal active inference that arbitrate between identity exploration and commitment. They propose that impaired self-modelling is unable to reduce interpersonal uncertainty during identity exploration, which, over time, degenerates into uncontrollable epistemic habits that isolate the body image from corrective sensory evidence. In light of these insights, the authors subsequently explore some of the consequences of image-centered social media platforms on the identity formation process. They conclude that heightened interpersonal uncertainty in this novel context could precipitate the onset of body dysmorphic disorder and related body image distortions, particularly when selfie filters are involved. With the progressive distortion of the body image through visual normalization, the automatization of epistemic behaviors is particularly problematic. Without the ability to successfully interrupt these uninformative epistemic behaviors, the distorted body image is isolated from corrective sensory evidence. Discussing the identity formation process in the online environment, the authors highlight its potentially damaging effects on interpersonal uncertainty minimization. Image-centered social media platforms could increase the pressure on this complex developmental task, increasing the probability of superficial identity commitments centered on appearance. This could explain the increasing prevalence of body image distortion phenomena that seem to implicate AI-powered filter-based social media apps.

F. LeRon Shults et al., in 'Minding Morality' (this volume), discuss the impact of policy-oriented computer models on policy assessments that ignore crucial social contextual factors, such as distinctive moral and normative dimensions of cultural contexts. They asserts that the incorporation of morally salient dimensions of a culture is critically important for producing relevant and accurate evaluations of social policy when using multi-agent artificial intelligence models and simulations. They conclude that social norms, and more generally the moral and ethical dimensions of human social life, are more than optional considerations for computational social scientists; they are critical for the relevance of policy simulation.

Edin Šabić et al., in 'Healthcare and Anomaly Detection' (this volume), discus the application of machine learning algorithms to healthcare data for enhancing patient care while also reducing healthcare worker cognitive load. They conclude that simulated data can help tune algorithms to some degree of performance when real labeled data is unavailable, and this type of imposed rule-based learning might be especially helpful when initially employing a system without any prior data.

Bokolo Anthony Jnr, in 'A Case Based Reasoning Recommender System for Sustainable Smart City Development' (this volume), discusses the need for stakeholders to make strategic decisions on how to implement smart city initiatives. He notes that currently city planners/developers are faced with inadequate contextual information on the dimensions of smart city required to achieve a sustainable society. Whilst recognizing the application of methods such as big data, Internet of Things (IoT), cloud computing, to support smart city attainment, the author argues for the integration of Case Based Reasoning (CBR) technique to develop a recommender system towards promoting smart city planning.

Roger Andre Søraa et al., in 'Children's perceptions of social robots' (this volume), study the perceptual differences of three social robots by elementary school children at the Norwegian national research fair. By comparing three different types of social robots, the study found that *presence* can be differently understood and conceptualized with different robots, especially relating to their function and "aliveness." The authors conclude that there exists a strong difference when relating robots to personal relations to one's own grandparents, versus the elderly in general. The article finds that children's perceptions of robots tend to be positive, curious and exploratory. These perceptions are possibly more guided by an emotional connection, rather than a rational interpretation based on culture and notions of usefulness when compared with those experienced by adults.

Jurriaan van Diggelen et al., in 'Hybrid Collective Intelligence in a Human-AI Society' (this volume), present a discussion on the future impact of Artificial Intelligence (AI) on human society, from three different perspectives: (1) the *technology-centric perspective*, claiming that AI will soon outperform humankind in all areas, and that the primary threat for humankind is superintelligence; (2) the *human-centric perspective*, claiming that humans will always remain superior to AI when it comes to social and societal aspects, and that the main threat of AI is that humankind's social nature is overlooked in technological designs; and (3) *the collective intelligence-centric perspective*, claiming that true intelligence lies in the collective of intelligent agents, both human and artificial, and that the main threat for humankind is that technological designs create problems at the collective, systemic level that are hard to oversee and control. The authors conclude that each of the three perspectives offers a unique contribution to the debate resulting from their differences in focus and background knowledge in specific applications and corresponding opportunities, risks, and challenges. It is further argued that combining the three perspectives into a single integrated and comprehensive framework allows for researchers and developers to adopt an appropriate perspective when tackling a given design challenge.

Monika Simmler and Ruth Frischknecht, in 'A Taxonomy of Human–Machine Collaboration' (this volume), discuss the challenges of ongoing advancements in technology for governance and accountability, especially that of increasing prevalence of the socio-technical collaboration. It is argued that it is thus crucial to familiarize decision makers and researchers with the very core of human–machine collaboration. However, socio-technical constellations are complex, and capturing them adequately calls for a distillation to their basic characteristics involved in human–machine collaboration: automation and autonomy. The paper introduces a taxonomy that enables identification of the very nature of human–machine interaction. It is noted that this taxonomy allows the user to grasp complex phenomena and focus on what is most important in human–machine collaboration: the question of who does what on one hand, and the question of how independent it is done on the other hand. Such a distribution of agency will allow professionals and researchers from different fields to estimate and evaluate the implications and consequences of given socio-technical constellations. By introducing different levels of automation and autonomy, the authors say that these concepts are not a question of all or none, but rather vary in degree, allowing for multi-faceted collaboration, while also allowing for all users, despite their level of specialized knowledge, a pragmatic means of capturing them.

Xueliang (Sean) Li et al., in 'Things that Help Out' (this volume), propose an approach to designing smart wearables that act as partners to help people cope with stress in daily life. This approach, the authors assert, contributes to the developing field of smart wearables by addressing how technological capabilities can be designed to establish partnerships that consider the person, the situation, and the appropriate type of support. The authors present the results of a phenomenological study conducted with three war veterans who suffer from chronic posttraumatic stress disorder. They describe how their experiences of dealing with their stress informed their design approach, and discuss the implications of these results on smart wearables and stress management in general. The authors conclude by reflecting on the limitations of this study and directions for future work.

Tahmina Khan Tithi et al., in 'Context, Design and Conveyance of Information' (this volume), study the design of effective and sustainable information services for marginalized women farmers in developing countries, and make recommendations for domain of context-specific information system design in resolving socio-economic problems to pave the way of sustainable development. The study uses PROTIC (Participatory Research and Ownership with Technology, Information and Change), and argue that a well-designed ICT4D solution must be tailored to the needs of the people. This requires an extensive understanding of the context and constraints in people's lives.

Javier Camacho Ibáñez et al., in 'Moral control and ownership in AI systems' (this volume), discuss how and under what circumstances the 'human subject' may, totally or partially, lose moral control and ownership over AI technologies. They note that AI systems can be designed to leave moral control in human hands, to obstruct or diminish that moral control, or even to prevent it, replacing human morality with pre-packaged or developed 'solutions' by the 'intelligent' machine itself. However, the issue of embedding moral agency in the machine becomes much more problematic when we consider complex networks of machines-eventually very different-that feed information and decisions into each other and to human operators. They argue that as moral agency forms a basis of our system of legal responsibility, complex AI networks become essential for the functioning of our societies, and the preservation of moral agency through them acquires bigger relevance.

Sergey B. Kulikov in 'Artificial Intelligence, Culture and Education' (this volume), citing Lotman and Uspensky, expands the meaning of artificial intelligence, and identifies a cultural type of 'strong' artificial intelligence or 'self-increase of *Logos*'. For the author, there is an equivalence of culture as a self-organized sign system and 'strong' artificial intelligence. Culture in this view produces signs and symbols regardless of possible agents external to it. Semiotics thus ensures the description of self-organizing systems of cultural signs and symbols in terms of artificial intelligence as a special set of algorithms. These signs and symbols can be systematically used in education. Autonomy of actions thus makes it possible to connect culture and artificial intelligence. Here, autonomy corresponds to the automatic formation of positive emotions and social orientations. From the empirical studies, the author posits that the organization of collective activities without external control ensures the development of positive emotions and social orientations. Interest in autonomous behavior provides the formation of educational and cognitive motives. As a special set of algorithms, the author argues that these motives are the most promising and favorable for personal development. The author finds that the boundaries between natural and non-natural forms of intelligent activity coincide with the sphere of behavior and communication. In this regard, artificial intelligence is the imitation of human mind.

Martin Miernicki and Irene Ng, in 'Artificial Intelligence and Moral Rights' (this volume), discuss the issue of moral rights in connection with moral rights and content produced by artificial intelligence, in particular whether an artificial intelligence itself, or the creators or users of an artificial intelligence should be considered as owners of moral rights. They note that most of the discussion on moral rights focuses on economic rights and protecting the author's "personal sphere", whereas the relationship of artificial intelligence and moral rights remains relatively obscure. While right to integrity is perhaps fundamentally rooted in the personal sphere of the author, the attribution can also be explained on the basis of other foundations. In this connection, intermediate solutions are conceivable, such as the introduction of "contributorship" right. It is recognized that there are practical problems such as establishing whether the work is indeed the product of human creative endeavors and associated evidentiary problems. One possible strand of thought in favor of acknowledging moral rights is that these rights serve a higher public function—attributing moral rights to the original creators (who arguably have the greatest interest in protecting their own works) could not only be understood as serving the authors' own interest but also the public interest in the form of integrity and societal status of creative works, in general, for the greater good.

Daniel Schiff, in 'Out of the Laboratory and into the Classroom' (this volume), discusses AI's impacts on education policy and practice, paying special attention to intelligent tutoring systems and anthropomorphized artificial educational agents. The discussion includes the role of stakeholders towards improving their engagement with socially responsible research and implementation of AI in educational systems. Whilst recognizing the disruptive impact of artificial intelligence, the discussion also recognizes the potential of peer review as a gatekeeping strategy to prevent harmful impact of AI for education research.

Inbar Kaminsky, in 'Do Robots Dream of Escaping?' (this volume), addresses the dilemma of whether or not the humanoids in films, *Ex-Machina* and *Morgan*, possess high levels of artificial consciousness and consequences of focalization. The author further addresses ethical issues that are raised throughout the films in relation to confinement and surveillance, and argues that the 'hard question' of consciousness becomes even harder when dealing with its possible emergence among humanoids. If humanoids are indeed endowed with a subjective inner life, then they are entitled to be treated as moral agents, equivalent to humans rather than animals.

Tomáš Zemčík, in 'Failure of Chatbot Tay' (this volume), discusses the anthropomorphization of chatbot Tay, and explores whether communication of an algorithm with society using socio-centric or individualistic morals can have a significant impact. The article notes that 'there is a certain kind of cognitive distortion, where people know that they do not communicate with a human being (contrary to the Turing test) but they still ascribe higher level of existence (intelligence, responsibility, soulfulness and others) to chatbot's comments, agency and its whole 'personality' than the chatbot objectively possesses'.

Kwame Porter Robinson et al., in 'Authente-Kente' (this volume), propose the development of a cell phone based authentication app for 'kente' cloth in west Africa, and discuss an initial test of a *machine learning* algorithm for

distinguishing between real and fake 'kente', emphasizing the importance of a larger social-technical context for the next stages of development. It is important to note that such platforms would focus on connecting local artisans and regional data for continuous training as a way to improve accuracy, or selectively account for prevalence to improve overall decision reliability. By design this platform makes no assumption and assumes the prevalence rate is equal. This, the authors say, can be addressed in future work with the mobile application that contains the machine learning pipeline.

Sadia Sharmin and Danial Chakma, in 'Attention-based Convolutional Neural Network for Bangla Sentiment Analysis' (this volume), present how the attention mechanism could be incorporated effectively and efficiently in analyzing the Bangla sentiment or opinion. They note that their empirical study about Bangla sentiment analysis is a small step forward to fill the void in both the benchmark datasets and a well-furnished model for sentiment analysis for Bangla, despite being one of the most used languages in the world.

Stephen Edwards, in 'AI in the Noosphere' reminds us of the solace and freedom of spirit we find in the healing meditation and related reflections. We learn about the 'Freedom to be what we are and to become what we can become' as the guiding philosophy for human sustainability when going through tough times or constraining conditions. The author argues for an alignment of most scientific and wisdom traditions with everyday practice as power tools of the heart such as appreciation, forgiveness, non-judgement, peacefulness, care and love.

Luciano Floridi et al., in 'The Chinese Approach to Artificial Intelligence' (this volume), discuss the impact of China's AI strategy on three areas of policy: *international competitiveness*, *economic growth*, and *social governance* (*construction*). The authors argue that AI can help foster increased productivity and high levels of growth, but its use is likely to intensify the inequalities present within society and even decrease support for the government and its policies. Commenting on the impact of AI strategy in the areas of privacy and medical ethics, they identify many loopholes and exceptions that enable the government (and companies implicitly endorsed by the government) to bypass privacy protection and fundamental issues concerning lack of accountability and government's unrestrained decisional power about mass-surveillance. In the same vein, when focusing on medical ethics, it is clear that, although China may agree with the West on the bioethical principles, its focus on the health of the population, in contrast to the West's focus on the health of the individual, may easily lead to unethical outcomes (the sacrifice imposed on one for the benefit of many) and thus creating a number of risks, as AI encroaches on the medical space. These are likely to evolve over time, but the risks of unequal care between those who can afford a human clinician and those who cannot, control of social diseases, and of unethical medical research are currently the most significant. The authors posit that it is also important to understand all this not just externally, from a Western perspective, but also internally, from a Chinese perspective.

Nik Hynek and Anzhelika Solovyeva, in 'Operations of Power in Autonomous Weapon Systems (this volume)', provide a multi-perspective examination of one of the most important contemporary security issues: weaponized, and especially lethal, artificial intelligence. They note that this technology is increasingly associated with the approaching dramatic change in the nature of warfare. What becomes particularly important and evermore intensely contested is how it becomes embedded with and concurrently impacts two social structures: ethics and law. While there has not been a global regime banning this technology, regulatory attempts at establishing a ban have intensified along with acts of resistance and blocking coalitions. The authors reflect on the prospects and limitations, as well as on the ethical *and* legal intensity of the emerging regulatory framework.

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

## Reference

Gill KS (2020) Ethics of engagement. AI Soc 35(4):783–793

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