AI & SOCIETY (2021) 36:669–676
https://doi.org/10.1007/s00146-021-01260-7

EDITORIAL

Ethical dilemmas

Ned Ludd and the ethical machine

Karamjit S. Gill[1]

Accepted: 11 August 2021 / Published online: 17 August 2021

Smith (2021) in 'Perhaps Ned Ludd had a point?', alerts us to pay attention not only to the great ethical and theoretical issues, but also to the ways in which actual human lives are impacted by what we think, say and do. We should be vigilant of a widespread tendency to both overestimate the pace and underestimate the extent of change of advanced technologies in a variety of real-world contexts. It may be tempting to be consumed by a positivistic and empirical world view that focuses on 'how humans judge machines' and not on 'how humans could and should judge machines' (Gill 2019). The human-centred ethos of AI&Society asks us to transcend this techno-centric view, and explore not just the 'how' question, but also the 'could' and 'should' questions. In exploring these questions, we need to be mindful of the concern that, for example, algorithmic aversion may risk rejecting technology that could improve social welfare and we may 'fail to recognise the consequences of technology when we show a positive bias towards algorithms.' (Gill 2020, 2021). Whilst the techno-centric paradigm tends to provide efficiency, precision and replicability of technological innovations, the human-centred paradigm promotes creativity, flexibility, and resilience. Those who seek the trade-off between efficiency and flexibility face ethical challenges that designers of all technologies face.

AI&Society authors continue to reflect on narratives of AI ethics that vary from moral and ethical dilemmas of human judgment in the 'heat of the moment' of the trolley problem, to ethical implications such as those of opacity, explainability, reliability, trustworthiness and justice that arise from the development and implementation of artificial intelligence (AI) technologies. Self-driving cars open 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. The narrative on moral machines and 'virtuous ethics' gives an insight into the relational functions of social robots such that of providing empathy and intimacy or even encouragement and advice. From this perspective, moral machines 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. The argument is that what matters ultimately is the flourishing of the virtuous agent, and virtue's benefits for society such as trustworthiness, safety, etc. If so, then the virtues in question are only instrumental. It is argued that even in this case, we *encounter* virtuous agents 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. The use of predictive systems in socially and politically sensitive areas such as crime prevention and justice management, and crowd management and emotion analysis, raise ethical concerns of misclassification, for example in the case of conviction risk assessment or the decision-making process, when designing public policies (Gill 2020, 2021). It is argued that such automated AI decision support systems might perpetuate bias that is already in the data used to set up the system, e.g., by increasing police patrols in an area and discovering more crime in that area. Although there is a general discussion about privacy and surveillance in information technology, focusing mainly on the access to private data and data that is personally identifiable, the ethical narrative of AI in surveillance goes beyond the mere *accumulation* of data and direction of attention: they include the *use* of information to manipulate behaviour, online and offline, in a way that undermines autonomous

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

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

rational choice. Opacity and bias are central issues in what is now sometimes called 'data ethics' or 'big data ethics'.

The design of the algorithmic agency for societal applications prompts questions about regulatory issues of, for example, 'accountability', 'responsibility', and 'liability' and normative issues of, for example, 'fairness', 'informed consent' and 'avoiding bias', and technical questions of 'interoperability' and 'updatability', and organizational issues of, for example, 'feasibility, decision-making and 'interventions'. Any ethical narrative exploring such issues, needs to take account of different contexts in which societal interactions take place. We may thus argue that the incorporation of morally salient dimensions, social and cultural, is critically important for producing relevant and accurate evaluations of social policy, when using multi-agent artificial intelligence tools. Furthermore, the challenge is whether we can gain enough understanding of the processes inherent to ethical decisions, to the point that these can be 'taught' to the machine for agency capable of manifesting ethical discernment, especially as machines become active players in societal dimensions that, until now, have been attributed exclusively to humans.

The European Parliament Report (2020) explores ethical narratives with broader societal contexts. Examples of ethical narratives include the impact of AI on relationships, as in the case of intelligent robots taking on human social roles, such as nursing, affecting human–human relationships in as yet unanticipated ways. These include ethical concerns in relation to the deployment of robots for the care of the elderly in particular. The use of AI in healthcare also raises questions about trust, for example, how trust in professionals might change if they are seen as 'users' of technology. Self-driving autonomous cars are likely to raise issues of liability, trust, human respect and civil rights. Autonomous weapons and drone technologies raise ethical issues of human judgement. From a societal perspective, the ethical narrative moves beyond the technological design to issues of human rights and well-being; emotional harm; accountability and responsibility; security, privacy, accessibility and transparency; safety and trust; social harm and social justice; lawfulness and justice; control and the ethical use (or misuse) of AI; environmental harm and sustainability; informed use; existential risk.

However, innovation of the AI agency poses a challenge of creating a culture of responsible innovation that involves 'the task of building an accessible moral vocabulary' for ethical engagement. Leslie (2019) says that such a moral vocabulary draws primarily on two traditions of moral thinking: (1) bioethics and (2) human rights discourse. Whilst bioethics is concerned with the study of the ethical impacts of biomedicine and the applied life sciences, human rights discourse draws inspiration from the UN Declaration of Human Rights. It is anchored in a set of universal principles that build upon the idea that all humans have an equal moral status as bearers of intrinsic human dignity. Whereas Bioethics largely stresses the normative values that underlie the safeguarding of individuals in instances, where technological practices affect their interests and wellbeing, Human Rights discourse mainly focuses on the set of social, political, and legal entitlements that are due to all human beings under a universal framework of juridicial protection and the rule of law. According to Leslie (ibid.), the main principles of bioethics include respecting the autonomy of the individual, protecting people from harm, looking after the well-being of others, and treating all individuals equitably and justly. The main tenets of human rights include the entitlement to equal freedom and dignity under the law, the protection of civil, political, and social rights, the universal recognition of personhood, and the right to free and unencumbered participation in the life of the community. This discussion raises further questions about the driving forces of AI systems and their impact on shaping the future of data-driven society, including their influence on identity transformation, and how these forces influence our purposes and values as morally and socially responsible human beings.

From a human-centred perspective, the challenge facing the data driven society is how to keep the human-in-the-loop and shape AI systems that create a culture of ethics and enhance symbiotic collaborations between humans and machines. This raises questions of responsibility and intentionality, and of 'legitimacy' and acceptability. And further, 'what and how' ethical, moral or aesthetic choices are not to be made by the AI machine but by those who use them as ethical, moral or aesthetic agents in domains of governance or autonomous decision-making. In exploring the issue of AI governance, we may ask whether a utilitarian-centred machine can accommodate various societal needs, and whether socio-technical systems can fill the gap between normative and moral ethics. And further, whether we can proceed from the 'descriptive level to the explanatory level, then to the level of interventional ethics and towards reconciliation and reformation'.

Many of our journal authors explore ethical implications and consequences of the AI machine beyond theoretical ethics. For example, authors in this volume discuss the effectiveness of the electronic portal as a mediated communication tool that supports the tacit engagement of mental health carers, thereby serving as a virtual bridge for therapeutic communication and addressing mental health isolation not as a medical issue but as a social one of mental health therapy, a humanised experience of recovery, in recognition of both the transparency and non-transparency for promoting mental health recovery. It is recognised that there is a need to formulate an ethical framework for the machine learning community that is engaged in human–computer interaction design that caters for the real needs of the health care

users including patient's informed consent and obligation of medical professionals to warn them about potentially harmful consequences of diagnosis. From a broader societal perspective, such an ethical framework should encompass the general obligation to warn the users of possible misuse of information dissemination and harmful consequences of malicious use of machine learning applications, and risks of misusing software as a tool to harm or suppress other people. These explorations are underpinned by a discussion on the ethical choices we can have and the degree of autonomy that should be given to the building of machine invention systems, raising issues such that of morally driven approaches to AI ethics and responsibility, moral divide of 'human vs. nonhuman', and building ethics into machines. This also applies to the design of social machines that place value in human interaction and act as ethical support tools for ethical and technological discourses. The design of such tools confronts the ethical and policy dilemmas that not every moral decision should be outsourced to machines. We are alerted to the limit of 'techno-evidence' to resolve the issue of trust and growing distrust of machines in an increasingly complex society, with concerns over reproducing societal bias, especially the notion of precision in the quality of evidence, that may well be at odds with the values that underlie the societal processes. We should always remember that the goal of making the world a better or fairer place requires a great deal more than ethical frameworks for data science, it requires cultivation of human machine collaborative engagement that enriches and enhances actual human lives.

Robert Sparrow, in 'Why machines can't be moral' (this volume), provides an insight into the flaw of building "ethics" "into" machines, arising from the presupposition of a flawed understanding of the nature of ethics. The author argues that machines could at best be engineered to provide a shallow simulacrum of ethics, which would have limited utility in confronting ethical and policy dilemmas associated with AI. On the nature of the ethical, the author posits that it is constituted by our practices of moral reasoning, and suggests that only creatures with bodies and faces with the expressive capacities of—if not identical to—those of human beings, can be justified in saying that they can experience the ethical. In acknowledging ethical dilemmas from the objective and of explanatory perspectives, the author concludes that before we try to build ethics into machines, we should ensure that we understand ethics. Anuradha Reddy et al., in 'Encountering ethics *through* design' (this volume), provide an insight into whether intelligent things can have an ethical agenda, and if so, could we then imagine ways to move past the moral divide 'human vs. nonhuman' in those contexts, where things act on our behalf? The insight arises from a scenario workshop on intelligent things that allows unforeseen ethical situations to emerge in an improvisatory manner. We learn that by giving intelligent things an active role in interaction, the workshop participants seemed to be activated by the artifacts, provoked to act and respond to things beyond the artifact itself—its direct functionality and user experience. It is suggested that the workshop helped to consider autonomous behaviour not as a simplistic exercise of anthropomorphization, but within the more significant ecosystems of relations, practices and values of which intelligent things are a part. The authors suggest that the workshop can be seen to contribute to and complement morally driven approaches to AI ethics and responsibility. This allows participants to think creatively about the effects of interacting with intelligent things in everyday life and the implications that these interactions bear on society. Formosa and Ryan, in 'Making Moral Machines' (this volume), first ask whether we should seek to create Artificial Moral Agents (AMAs), and then argue that all things considered we have strong reasons to continue to responsibly develop AMAs. The authors note that not every machine should become an AMA and not every moral decision should be outsourced to machines. They stress the careful use of AMAs in sensitive contexts, taking account of issues around responsibility and trust. Sekiguchi and Hori, in 'Designing ethical artifacts has resulted in creative design' (this volume), discuss the design of an ethical support tool to improve the creativity of an engineer's design activity. Designed around the application of an ethical design theory, the support tool provides an environment for the promotion of ethical design perspectives and description. The proposed ethical design theory extends the hierarchical representation of artifacts, thereby enabling users to reconsider their themes at the highest level of the hierarchy and apply a wider conceptual space of design solutions. It is noted that both these functions are realized by exploiting a knowledge base of ethical and technological discourses. Based on their study of ethical design, the authors further note that the ethical design theory can be updated based on some unexpected results with regard to the cyclic relationship among theory, tools (i.e., experimental equipment), and observed data. The authors suggest that using the scenario path recommendation, designers can update their research themes after considering the ethical impacts of those themes on stakeholders.

Bruneault and Laflamme, in 'AI Ethics: How Can Information Ethics Provide a Framework to Avoid Usual Conceptual Pitfalls?' (this volume), note that whilst there is a considerable research interest in artificial intelligence ethics (AIE), the focus primarily remains on ethical issues specific to certain areas of expertise, thereby this focus often remains confined to narrow areas of application, without considering the global ethical issues in which they are embedded. The authors discuss an alternative approach of informational ethics that takes into account the political issues that emerge from the social deployment of AI. Ratti and Bezuidenhout, in 'What Does It Mean To Embed Ethics In Data Science?'

(this volume), discuss how ethics should be embedded in the practice of data science, in the sense of showing how ethical issues emerge in small technical choices made by data scientists in their day-to-day activities, and how such an approach can be used to teach data ethics. The authors propose the use of emerging models of 'micro-ethics', as a tool for teaching daily responsibility in digital activities that is connected to (and draws from) the higher level ethical challenges discussed in digital/data ethics. It is, however, recognized that the applicability of this tool is reflected also in the way data ethics is taught, especially data ethics to data scientists. Furthermore, it is recognized that stand-alone courses based on macro-ethical issues struggle to make a direct connection between ethical issues and daily practice of data science, and thus would benefit from grounding teaching strategies within a virtue ethics framework.

Moa De Lucia Dahlbeck, in 'AI and Spinoza' (this volume), suggests that Spinoza's philosophy of mind and knowledge may function as an analytical tool for making sense of the prevailing conception of AI within the legal discourse on Lethal Autonomous Weapons Systems (LAWS). To make sense of the problem of AI in law, the author first contemplates ethical and political discussions of intelligence and human cognition together, so as to identify separate but inter-related grains of critique of the law-making process taking place under the auspices of the CCW negotiations on a LAWS protocol. It is argued that in the light of Spinoza's normative theory of judgment, the fear and hope generated in human beings by their affective encounter with AI is more dangerous and detrimental for peace and stability than AI taken on its own. In the light of this, the author proposes that the legal discussion on how to regulate human interactions with AI must perhaps endorse adaptable and varying legal measures and norms, according to the specific desires and impulses that dominate within different particular contexts in which they are to function.

Jebari and Lundborg, in 'Artificial superintelligence and its limits: why AlphaZero cannot become a general agent' (this volume), explore the characteristics of machine agency, and what it would mean for a machine to become a general agent. The authors argue that to become a general agent, a machine needs *productive* desires, or desires that can direct behaviour across multiple contexts. However, productive desires cannot *sui generis* be derived from non-productive desires. Thus, even though a general agency in AI could, in principle, be created by human agents, the general agency cannot be spontaneously produced by a non-general AI agent through an endogenous process (i.e., self-improvement). In conclusion, the argument is that a common AI scenario, where general agency suddenly emerges in a non-general agent AI, such as DeepMind's superintelligent board game AI AlphaZero, is not plausible. The paper concludes by noting that rather than being complacent about AI risk, measures to monitor and guide the development of AI are potentially feasible.

Daniel Innerarity, in 'Making The Black Box Society Transparent' (this volume), discusses the demands of transparency and non-transparency to reduce the ignorance of automated decision-making processes in our societies. The paper examines a promising concept of explainability by placing it in the framework of collective capacities to design a possible comprehensibility. The author suggests that we need to think about what kind of capabilities and collective intelligence would be needed to make automation compatible with the ideals of autonomy and responsibility in a human-centred technological environment.

Manuel Carabantes in 'The Internet as a Heideggerian Paradigm of Modern Technology' (this volume), views Internet as a paradigm of modern technology in the Heideggerian sense in that: First, it is a mode of revealing (*Entbergen*) that performs a setting-upon (*Stellen*). Second, it is a challenging (*Herausfordern*) revealing that violently demands the presence of what-is (*Seiende*) without waiting and without uncertainty, which is different from the revealing of traditional technology. Third, the standing-reserve (*Bestand*) is its relative mode of appearing of what-is, which means that what-is appears or presents (*anwest*) as available reserve to be exploited. And fourth, it produces a multiple concealment (*Verborgenheit*) that also highlights the concealment of our own Being. From this perspective there is an argument against mythinformation philosophies that say that the Internet is not governed by an alleged non-dominant, dialogical, and cooperative operativity.

Thilo Hagendorff, in 'Forbidden knowledge in machine learning' (this volume), makes a case for transferring the discourse on 'forbidden knowledge' (too sensitive, dangerous or taboo), to machine learning research. The discussion recognizes the possible misuse of information dissemination and harmful consequences of malicious use of machine learning applications. It is, however, argued that the idea of forbidden knowledge in machine learning should not put limits or constraints on science or the pursuit of legitimate research questions—but should put limits on the way research insights are shared. These limits should be established not because machine learning science itself is dangerous. Rather, it is the current political and cultural climate in many parts of the world that brings forth risks of misusing software as a tool to harm or suppress other people. The author proposes a tentative ethical framework for the machine learning community on how to deal with forbidden knowledge and dual-use applications.

Vinícius P. Gonçalves et al., in 'FlexPersonas' (this volume), set out a collaborative method called FlexPersonas. It is employed for the flexible mapping of health care users with a view to improving decision-making with the support of Internet of Things technologies. In this context,

computing systems can identify the behavior of the user, and issue warnings to the carers of rehabilitation treatment and senior citizens about any abnormal event (for example falls or accidents). The authors note that as technology, interfaces, artifacts and even the users, are continually evolving, researchers face new challenges of human–computer interaction design to meet the real needs of health care users including human factors when developing technologies.

Karuna et al. in 'IoT Plant Monitoring System for Mental Health Therapy' (this volume), propose an Internet of Things (IoT) tool to enhance the experience of personal gardening as a method of therapy for mental-health patients, given a belief in its role in a person's mood and general positivity. The authors propose an (IoT) prototype that continuously senses and monitors the state of an indoor plant through different sensors. In this prototype, the user is notified of the plant's needs for water, sunlight, through generated notifications from channels over 'cloud' in-real time. It is noted that the creation of a smartphone mediation provides for a humanised experience of recovery. It is recognised that mental health recovery often revolves around therapeutic exercises that may contribute to one's personal development across a journey of healing.

Maximilian Kiener in 'Artificial Intelligence in Medicine and the Disclosure of Risks' (this volume), introduces the reader to medical disclosure in clinical settings that arises from the increasing applications of AI in medicine. It asks whether the physician needs to disclose AI risks to patients, i.e., the risk of a cyber-attack, the risk of bias affecting a patient's health care, and the risk of a mismatch. The author argues that, under certain circumstances, these risks do need to be disclosed. Otherwise, the physician either vitiates a patient's informed consent or violates a more general obligation to warn him about potentially harmful consequences, especially when these risks are exacerbated by pandemics like the COVID-19 crisis.

Illankoon and Tretten, in 'Collaborating AI and Human Experts in the Maintenance Domain' (this volume), discus the need for a better understanding of the linkage between the technicians' knowledge and Intelligent Decision Support Systems. This linkage builds upon the dynamics between distribution of knowledge among different agents, and collaboration of knowledge for reaching a shared goal, whilst recognizing the dynamic challenges involved in operational level maintenance. It is posited that since the technology of Augmented Reality (AR) uses both distribution and collaboration concepts, the paper recommends AR based maintenance decision support systems.

Lees et al., in 'IIoT and Cyber-Resilience' (this volume) discuss cyber threats and risks of automated and networked transnational supply chains for society in areas including business, environment and health. Although blockchain capabilities and machine learning technologies can be introduced to improve the technical and organizational operations, the advancement and proliferation of IIoT has increased the attack surface and vulnerability of the contemporary enterprise. It is suggested that although existing cyber security protection methods are arguably inadequate for managing the risks in the emerging digital world, they are often not implemented as designed and hence fail to achieve full benefit.

Zhitomirsky-Geffet and Weic in 'Utilizing Facebook for professional integration of three ethnic groups in Israel' (this volume), study the influence of social network behaviour of different ethnic groups, and the role of social networking sites as a catalyst for the creation of intergroup professional relations. It is proposed that the utilization of social networking sites as a platform for professional promotion might constitute a first step in the process of professional and cultural integration of minorities in the ethnically heterogeneous society. The study proposes a conceptual model for utilization of Facebook for professional integration of ethnic minorities, based on the social capital and weak social ties theories. It is argued that the proposed model provides an effective tool for estimating the level of professional integration of minorities, thereby increasing their willingness to create intergroup relationships that might lead to expanding professional circles and enhancing professional integration of minorities.

Hebblewhite and Gillett, in 'Every Step You Take, We'll Be Watching You' (this volume), discuss the way the increased usage of GPS devices is having a significant impact on human neuro-cognitive systems, especially memory and perception. They explore how habitual reliance on GPS technology undermines autonomous decision-making through 'nudging' in the sense of the alteration of psychological behaviour without the explicit forbidding of choice. It is suggested that whilst the wayfinding GPS technologies may free us from the burden of a tedious cognitive task, they also sculpt the way we tackle these problems such that we only build the thinnest of cognitive maps of our environments. The very tools for navigability that are offered to us implicitly limit our ability to make choices by shaping the very way in which we navigate our environments, potentially making some choices imperceptible.

Tyler L. Jaynes, in 'Citizenship as the Exception to the Rule', discusses the impacts popular media have on imprinting notions of computerised behaviour and its subsequent consequences on the attribution of legal protections to AIS and on speculative technological advancement that would aid the sophistication of AIS. The author suggests that we must address the difficult questions facing our societies as to how sophisticated machine intelligence (MI) systems *ought* to be treated. This call for action should hold special weight considering the influence MI systems may gain once the pandemic has been abated internationally. Consequently

what may now be considered 'remote' or virtual work may be delegated to self-learning computer entities.

Vicari and Gaspari, in 'Analysis of news sentiments using Natural Language Processing and Deep Learning' (this volume), discuss the development of Deep Learning models as a tool for forecasting the market sentiment using news headlines, with a view to developing an algorithmic trading strategy and testing it in real-world scenarios. However, the authors acknowledge a potential danger, due to the proneness of such a strategy towards the 'herd behavior', for the financial system and thus must be handled carefully.

Petar Radanliev et al., in 'Artificial intelligence in cyber physical systems' (this volume) note an increased attention to IoT and its juxtaposition to other related systems and technologies (e.g., Industrial Internet of Things, Cyber Physical Systems, Industry 4.0 etc.). In offering an analysis of the evolution of AI decision-making in cyber physical systems, the authors suggest that this evolution is inevitable and autonomous because of the increased integration of connected devices (IoT) in cyber physical systems. It is proposed that complex interconnected and coupled cyber-physical systems (CPS) can evolve automatically with the continuous technological upgrades in existing CPS. Here these systems are perceived as social machines that place value in human interaction with such systems. The authors argue that there is a value for artificial intelligence to learn from human–computer interactions. Instead of relying only on feedback from connected devices, in some scenarios, human input is of much greater value. For example, COVID-19 contact tracing apps are based on human–computer input to overcome just the computer data input that was considered too slow and ineffective.

Vasilescu and Filzmoser, in 'Machine invention systems' (this volume), discuss current developments in fields such as quantum physics, fine arts, robotics or defense and security, indicating the emergence of machine invention systems that are capable of producing new and innovative solutions through combinations of machine learning algorithms. It is suggested that because of the revolutionizing potential of such machine invention systems, there are widespread implications to consider from ethical and moral implications to policymaking and societal changes. The authors posit the need for further development of a theoretical framework encompassing machine invention systems, to better understand the boundaries, capabilities, and limitations of the current state of the art.

Queiroz et al., in 'AI from Concrete to Abstract' (this volume), discus the importance of initiatives that help the general public to build a basic understanding of the future of artificial intelligence, and the choices they can have on ethical and autonomy that should be given to the building of intelligent systems. This article presents the conceptualization and design of a new methodology, AI from concrete to abstract (AIcon2abs), to endow general people (including children) with a minimum understanding of what AI means.

Parfett et al., in 'AI-based Healthcare: A New Dawn or Apartheid Revisited?' (this volume) discuss the potential for hidden 'prejudice', should Artificial Intelligence (AI) gain a dominant foothold in healthcare systems. Drawing upon the suffering of the Chinese population during the Bubonic Plague outbreak that wormed its way through San Francisco's Chinatown in 1900, the authors make us aware of the potential prejudice inherent in AI systems, from police prediction and facial recognition software to recruitment tools. We learn about the apparent human need to classify things and its potential prejudicial implications that come with the desire to sort people on group lines. Unless care is taken, the authors note, prejudices such as those that governed both San Francisco's Chinatown in 1900 and South Africa under apartheid will continue to emerge, only now they will emerge through AI systems. The authors remind us of the danger of 'Coded Gaze', when the views that are embedded into systems are propagated by those who have the power to code the systems. 'Whoever codes the system embeds her [their] views'.

Lode Lauwaert, in 'Artificial Intelligence and Ethics. Who's Responsible for a Robot's Mistakes?' (this volume), sheds some light on Sparrow's 'responsibility gap' and the condition for the admissibility of an act arising from the debate on moral arguments and the ban on Lethal Autonomous Weapons Systems (LAWS). The author concludes that Sparrow's justification for his claim that LAWS should be banned is insufficient, and neither we can conclude that the thesis of a responsibility gap has in any case been undermined.

Vladimir Tsyganov, in 'Socio-Political Stability, Voters Emotional Expectations, and Information Management' (this volume), examines the notion of dependence of socio-political stability on the emotional expectations of voters, from the perspective of 'Progressist society, and 'Phobic's society'. Taking an example of Eastern Europe, the author argues that socio-political stability environment is enough for the engagement of the Progressist society; but the Phobic's society needs regular support of containment and nursing of their phobias. It is suggested that in the absence of regular impacts supporting phobias, a Phobic turns into a Progressist. However, this socio-political system becomes unstable with a weak economy and growth limits. The author proposes that the contradiction between increased consumption and growth limits can be resolved using high humanitarian technologies without creating and using phobias. The paper concludes by saying that it is necessary to change the paradigm of unlimited growth of material consumption to the paradigm of non-material, spiritual development, if we were to aspire for the survival of humanity.

Bahadur Ibrahimov, in 'Intelligent Inspection Robotics' (this volume), discusses the deployment of 'Open Innovation' and identifies obstacles that arise from the entrenched organizations' traditions, values, and institutional culture. The author notes the need to resolve these obstacles and find solutions that fit into the cultural requirements of organisations, and ensure environmental and societal benefits to society. It is recognised that problems of industry are quite dense, and thus the innovation rate need to rise for the sake of the economy, environment, society, and humanity. The author proposes that more robotics and artificial intelligence should be implemented and adopted by industry, thereby embracing Open Innovation.

John McClellan Marshall, in 'TECHNOEVIDENCE: The "Turing Limit" 2020' (this volume), examines the oncoming socio-economic impact of the Technological Revolution and the 'AI Ecosystem', particularly on the legal community and its processes as a practical example with which both liberal artists and scientists might identify. The discussion also focuses on the tension between the economic and socially driven thrust of modern society and traditional human value systems. The article recognizes that the effects of the technological revolution on societal values and structures are likely to continue well into the foreseeable future, and suggests possible remedies for these problems avoiding the deep jargon of both the law and technology. We are alerted to the limit of 'technoevidence' to resolve the issue of trust and growing distrust of machines in an increasingly complex society. It draws our attention, especially, to the notion of precision in the quality of evidence that may well be at odds with the values that underlie the judicial process. The author thus argues that judges and lawyers who are inclined to depend upon 'technoevidence' as a crucial element to a judicial outcome may be ignoring the 'Turing limit' in favor of reaching an outcome, no matter how that outcome might conflict with the reality of human needs. The article concludes that the 'Turing limit' could be the dividing line in the judicial process between judges who experience the law and those who become bogged down in technique.

Ashkan Farhadi, in 'There is No "I" in "AI"', argues that self-awareness is a collaborative function of "I" and the mind. "I" is instrumental in the sense of self-awareness, but on its own, it is selfless. It is suggested that in addition, "I" is the heart of the decision-making process, and therefore, AI is missing "I", a selfless master of the mind. It is further suggested that whilst there is little doubt that artificial intelligence can gather and process information for reasoning in decision making, what sets our mind apart from AI is an entity independent from the mind called "I" that redeems our discretionary decision-making power independent of a deterministic principle of causality without the need for a metaphysical soul. It is posited that since "I" makes all the decisions for the mind, it represents a selfless master of mind and is the key element that is currently lacking in AI.

Mike Zajko, in 'Conservative AI and social inequality' (this volume) argues that concerns over reproducing societal bias should be informed by an understanding of the ways that inequality is continually reproduced in society—processes that AI systems are either complicit in, or can be designed to disrupt and counter. The discussion includes a contrast between conservative and radical approaches to AI, with conservatism referring to dominant tendencies that reproduce and strengthen the status quo, whilst radical approaches work to disrupt systemic forms of inequality. It is noted that given that politics is fundamentally about power, we would do well to recognize how these systems currently work to intensify, maintain, and optimize existing forms of power. It is also recognised that whilst interdisciplinary engagement can sometimes inform the design of AI systems to make them less harmful, the goal of making the world a better or fairer place requires a great deal more, especially when work in computing and data science continues to discriminate between social categories, without seriously engaging with what is known about these categories and their relationships in other disciplines.

Taebnia and Taqavi, in 'The Enhanced Human vs. The Virtuous Human', provide an insight into Farabi's concept of virtuousness of rational inquiry and deliberation. It is suggested that that this concept may be used to examine the virtuousness of the trajectories of enhancement technologies such as genetic engineering, neurostimulation technologies, or pharmacology. It is argued that although these technologies do not in themselves satisfy the constitutive determinants of virtuousness, they function as having both mediative and amplificative/reductive roles in a life, which is dedicated to the pursuit of happiness in the light of the cultivation of virtue.

As Larsson, in 'The wiseman in the mirror' (this volume) says that there are, however, other things to worry about. The application of AI and machine learning has unethically contributed to both reproducing existing inequalities in society and create realistic fakes and imitations.

Parviainen and Coeckelbergh, in 'The Political Choreography of the Sophia Robot', introduce the reader to the rhetoric about Sophia's citizenship, and move the discussion beyond recent discussions on the moral status or legal personhood of AI robots, to an analysis of the performativity of Sophia from the perspective of what the authors call 'political choreography'. It is this choreography that boosts the rise of the social robot market, rather than a statement about robot citizenship or artificial intelligence. Whilst criticizing the notion of 'embodied intelligence' used in the context of social robotics, the authors situate the discussions about the robot's rights or citizenship in the context of AI politics and economics.

Elissa Farrow, in 'Mindset Matters' (this volume), explores, through participant scenario workshops, the implication for organisational adaptation strategies when Artificial Intelligence (AI) is being embedded into the ecology of the organisation, and when employees have a dominant fixed or growth mindset. The author argues that growth mindset is a key component of adaptive capacity and literacy futures, and proposes five key components of this growth that include compassion and authenticity, embodiment as fundamental needs and motivations of mutual learning beyond the edges of the organisation. It is suggested that with an appropriate motivation, framing and mutual supportive environment, scenario workshop participants are able to switch from a fixed mindset, that was quite individualistic, to a growth mindset with more 'self actualised' societal or humanitarian considerations. The paper concludes by emphasising the need of an open creative and problem-solving mindset that can explore multiple scenarios, their complexity, and make decisions about how AI fits into new or impacted structural and societal models. It is suggested that bringing in the trans-contextual context of nature, family, community, biosphere and organisational, is critical to making sure we support our society to be resilient, aware and adaptive in shaping our futures.

Sue Pearson, in the Curmudgeon article, 'THE INSIDE OUT MIRROR' (this volume), asks whether we can free ourselves from our own automated algorithms and transform our external and internal worlds, whilst we seek to free ourselves from manipulation by tech companies. It is suggested that as automated algorithms in artificial intelligence can be seen as reflections of the brain's own algorithms, the online world can act as a mirror to reveal previously hidden internal workings of the brain that have impacted society for millennia. This inside-out mirror offers the possibility of transforming society itself. The author further makes the point that by turning the AI and internet mirror inwards, we have the insight to put the *reptilian core* in its place as a necessary survival part of the brain, but which has to be subservient to the thinking and decision-making of more evolved brain functions. This, the author says, would ensure people come before automated algorithms, and society comes before AI. However, first, we need to understand better how the brain and unconscious work.

Kimberly Cass in the Curmudgeon article, 'The Klein Bottle of Digital Identity' (this volume) reflects on 'What makes you "you?"—a perennial question of identity' in our digital age, whether the "you" is revealed by personal memories with others, 'inside-out' experiences, personal encounters and interactions, or unexpected apprehension, or 'real-time transitory glimpses that reveal who you are'. The author surmises that beyond our life in this world, the "you" is revealed by the thoughts and memories that are taken by and 'remembered when sparked by places, songs, words, environments, times of year. In many ways, the author concludes that: the "container" of our essence is formed by those who truly "know" us. "The song is ended, but the melody lingers on"—that elusive "something more" that remains in the hearts of those who have truly encountered "someone."

In its tradition of hospitality to diversity of argument and narrative, *AI&Society*, welcomes contributions on lived ethics, 'inside-out' experiences, societal encounters and interactions, and 'real-time transitory glimpses of the impact and implications of AI within societal contexts, in a way that not only keeps the human-centred song of AI&Society alive, but also nurture it across the horizon as melody linger on.

References

European Parliament (2020) The ethics of artificial intelligence: issues and initiatives. EPRS|European Parliamentary Research Service Scientific Foresight Unit (STOA)PE 634.452—March 2020. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf. Accessed 21 Apr 2021

Gill KS (2019) From judgment to calculation: the phenomenology of embodied skill. AI Soc 34:165–175. https://doi.org/10.1007/s00146-019-00884-0

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

Gill KS (2021) Ethical encounters. AI Soc 36(1):1–8

Leslie D (2019) Understanding artificial intelligence ethics and safety. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/201906/understanding_artificial_intelligence_ethics_and_safety.pdf. Accessed 23 Apr 2021

Smith D (2021) Perhaps Ned Ludd had a point? AI Soc. https://doi.org/10.1007/s00146-021-01172-6