AI & SOCIETY (2020) 35:1–4
https://doi.org/10.1007/s00146-019-00923-w

EDITORIAL

Dance of the artificial alignment and ethics

Karamjit S. Gill[1,2]

Published online: 14 November 2019

As we grapple with the impact of the artificial, on the one hand, and envision the common-good potential of augmented AI systems, on the other, we seek critical approaches to make sense of the nature of the emergence of data driven society. In this exploration, we face social challenges of governance, ethics, accountability and intervention arising from the accelerated integration of powerful artificial intelligence systems into core social institutions. We encounter the exponential rise of big data flows in networked communications and their manipulating algorithms, the gaps in translation are now too vast to grasp and address, rendering us unable to engage with difference through the shadows of machine thinking. Augmentation and automation places the human in the predicament to accept the calculation of the machine without judgment (Gill 2018). We echo Cooley's concerns of 'socially irresponsible' science (Cooley 2019) and ask whether we can transcend the instrumental reason of machine thinking (Weizenbaum 1976) to mould technological futures for common good rather than turning them into a single story of 'singularity'. Can we re-appropriate the idea of causality that has been taken by 'science' and reframe it in the making of everyday judgments and decisions? How can we harness collective intelligence as a transforming tool for addressing complex social problems? Just as narrators of the past situated their socio-technical debates in the context of their days, we need to draw upon various AI narratives of the relations between society and the scientific project of AI and the challenges it poses for us to come up with possible symbiotic AI futures. Just as the old technology arrived at a historical breaking point at which the old society was deemed to be transformed into a new one, the technologies of the artificial are now beginning to generate a situation in which society is once again facing the specter of a new transformation.

In exploring AI futures, we should take note of Cooley's (2019) reminder that the scientific project is always embedded within a particular social order and reflects the norms and ideology of that social order. In this perspective, science ceases to be seen as autonomous, as it internalises ideological assumptions thereby shaping the design of systems and tools and theoretical frameworks of its validation. It is undeniable that the drive of scientific knowledge has provided the material basis for a more full and dignified existence for the community as a whole, it must not however be a blind unthinking drive forward, shirking our social responsibility to critically examine its impact upon, and implications for, society.

In this issue, our authors contribute to an on-going debate on the narratives of artificial intelligence and society, ranging from alignment, digital hermeneutics, ethics, augmentation to global catastrophic risks. The alignment debate emphasises the alignment of the individual cognitive abilities to the collective in geographically distributed digital organizations, supported by the concept of human validation. Such an alignment builds upon the interplay between the tacit and cognitive dimensions of knowledge in creating participatory environments. The Curmudgeon wonders why is it that we hesitate to make causal inferences in favour of creative artificial agents and imaginative machines while we do not have such hesitation in drawing causal inferences in favour of creative human agents. Is it because of our species' chauvinism or anthropocentric bias? And how do we avoid being invidiously discriminatory in our responses to these imagination machines? We may not venture to answers such questions, but at least they may stimulate 'further lines of enquiry about imagination machines and their implications for our understanding of the imagination (creative or otherwise)'.

Here we encounter a symbiotic vision of human and machine alignment narrative, in which 'truly intelligent cognitive machines' perform 'human-like reasoning and learning' with the capability of 'human-like motivation, emotion,

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

[1] Professor Emeritus, Human Centred Systems, University of Brighton, Brighton, UK

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

and personality'. The argument for this idea of alignment is that it goes beyond the thesis of the computational intelligence of deep learning and reinforcement learning, thereby overcoming the narrowness cognitive paradigm in pursuing a fuller human–machine symbiosis. It is also argued that this debate on symbiosis can take advantage of the current momentum in the direction of the integration of the cognitive with social and natural sciences, thereby ameliorating the rising unease over the potential for AI and related technologies to shape the world going forward. The proponents of machine learning would, however, argue that 'statistical learning theory and machine learning models can be used to enhance understanding of AI-related epistemological issues regarding inductive reasoning and reliability of generalisations'. It is posited that this rather 'counterintuitive epistemological view of Deep Learning may provide an epistemological way forward and even perhaps an approach to how knowing is possible'.

From the idea of human–machine alignment we move towards an augmentation hypothesis of a collective human–machine subject of wearable technologies. In this perspective of digital hermeneutics, the argument is that as individuals in a community are augmented by the same wearable technologies, they 'generate a new type of collective subject, producing a collective organism in which individuals still pursue their collective goals'. In this vision, the 'world will become dwelled by these new collective subjects which will develop their own peculiar needs to be fed'.

When we envision the impact of AI systems in societal contexts, we are confronted with the propagation of big data either as a virtuous tool or as a malicious machine. However, we are reminded that 'contrary to the big data champions, big data is neither new nor a miracle without any error nor reliable and rigorous as assumed by its cheer leaders'. When seen from a societal context rather than as a technological advancement, the assumption of homogenous big data is rather misleading. It is thus crucial that society should resist the temptation of recognising big data as a virtuous tool of the internet of things. Moreover, the focus on technological advancement side steps the ethical dimension of system design, thereby creating a gap between the design and use of AI system and tools. In bridging the gaps in AI systems design and ethical dimension, it is posited that a practical way forwards lies in engaging 'AI engineers to grasp ethical issues by extending their own research and development, and practicing an ethical AI design' in the hermeneutics tradition. In developing the theme of digital hermeneutics, it may be asked in what ways digital technologies can legitimately be regarded as 'interpretational machines'. However, in asking this question, we are asked to reflect upon the history of interpretational beings-a 'second-degree reflection on the specificity of human beings as interpreting animals', thereby reflecting upon 'intrinsic difference between humans and digital machines (AI)'. In reflecting on the theme of digital hermeneutics, we note that the narrative of digital modernity may turn out to be a seductive concept, especially when seen in terms of the 'special affordances of digital networked technology'. It is, however, worth noting that this narrative tends to 'shape reality via commercial and political decision-makers'. Seen from an historical perspective, the idea of digital modernity may face space and time contradictions of its progression in the same way as contradictions that are found between ideas of digital modernity and modernity itself, and also between digital modernity and some of the basic pre-modern concepts that underlie the whole technology industry. It is posited that 'digital modernity may not therefore be a sustainable goal for technology development'.

As the technological vision of AI grapples with alignment or catastrophic risk, artists in the humanistic tradition of digital hermeneutics are beginning to show a way forward to designing AI systems and tools for engagement and collective intelligence. They do so through their inter-disciplinary and multi-disciplinary performances and installations. One such example is that of a Bird Song Diamond project of a multifaceted and multidisciplinary installation (this volume), in which engagement of public in the iterative design process provides 'artistic insights regarding the beneficial nature of collaborative interaction for promoting audience engagement with the subject matter'. This collaboration envisions that ethics resides in the human dimension and not in the machine dimension. From the humanistic perspective, the discussion on ethics of the artificial argues that AI designers should take responsibility for ethical design decisions, and these decisions must not be left for afterwards to AI systems, as we cannot "let the AI figure it out". It is noted that since we cannot take 'aggregate views of society' for granted, the choice between a social choice theory of ethics and a predetermined view of ethics is a rather weak argument for machine ethics. The ethical debate is thus seen as much broader than machine ethics. The argument is that ethics resides with the humans as decision makers and not in the 'behaviour of machines towards human users and perhaps other machines as well'. Since human ethics is centred around the idea of having a mind, the idea of equivalence of machine mind and human mind is very problematic. In this perspective, the 'very idea of an artificial moral agent or machine ethics fails to be a moral agent'. If the artificial agent fails the test of moral agency, then can we justify the idea of robots as evil machines, just as society does this for humans in certain unforgivable circumstances? It is interesting to note that whilst the idea of the evil machine as envisioned in robotic ethics research may lie in the argument that machines are evil because humans say so, there is no such tendency to 'depict intelligent machines as malicious', when we consider the issue of autonomy, privacy and liability. In thinking of the issues of autonomy, privacy and liability within the context of AI and Law, we are reminded that legal structures could act as barriers or enablers, depending on the adoption of automated systems. This is also the case with autonomous agricultural robots, for example the concerns of technology as an enabler or a barrier can be seen in 'the use of shared communications resources and privacy in the reuse of robot-collected data'. This raises the question of not just defining privacy but also of the scope of the autonomy of the robot. In arguing the case for designing intelligent machines that align with human values, we also raise the issue 'whether such highly advanced yet artificially intelligent beings will deserve moral protection once they become capable of moral reasoning and decision making.' In doing so, we also raise the question of machine ethics, of whether the intelligent machine in the form of an autonomous robot deserves to be granted 'moral rights once they have become full ethical agents, i.e. subjects of morality'. The ethics of the artificial agent debate goes beyond roboethics and asks whether the behaviour of autonomous agents such as autonomous vehicles, lethal autonomous weapons, and automated financial trading systems can now be evaluated as artificial moral agents. This notion of moral evaluation raises questions of how should artificial moral agents make decisions, and whether 'moral theory' is better placed than machine ethics. It is argued in this volume that the rule-based utilitarian approach for guiding the virtuous artificial moral agent, captures 'the most important features of the virtue-theoretic approach while realizing additional significant benefits'. And further the 'utilitarian artificial moral agent incorporating both established moral rules and a utility calculator is especially well-suited for machine ethics'.

As our society is experiencing an ever-growing integration of the Internet into everyday lives, we face the problem 'Internet Addiction' (IA) that has emerged from the problematic and excessive Internet usage, which leads to the development of addictive cyber-behaviours, causing health and social problems. It is proposed that one way forward is to develop 'Internet-based IA Recovery Framework (IARF) which uses AI to closely observe, visualize and analyse patient's Internet usage behaviour for possible staged intervention'. The ethical issue then is how to control the design and use such smart Internet-based systems. In the pursuit of reducing the negative impact of the questionable AI, the discussion wonders whether AI systems programmed with a virtual consciousness and conscience would reduce AI threats via motivational control, and whether other threats such as the desire for AI—human socio-economic equality could prove detrimental. Seen through a redemptive lens, however, the AI revolution is seen to bring extensive medical benefits to society, for example 'deep-learning neural networks can extract important information for big data bases by screening millions of skin abnormalities to diagnose a patient's abnormality, as well as find associations between a patient's condition and those patient's genetical, medical, physical, environmental and social records, in order to find the cause of Alzheimer'. In the same vein as the Internet addiction, this volume notes that 'phubbing' as a new phenomenon of ignoring conversational partners leads to diminishing and often producing negative interpersonal relations. This phenomenon raises concern of whether 'phubbing' gives rise to behaviour that is not only unaware of the social milleu, but also entails adverse effects and hurtful behavior. It is hypothesized that 'phubbing' develops the 'notion of *digital akrasia*, which can be defined as a tendency to become swept up by ones digital devices in spite of better intentions'.

As the discussion moves beyond the technological vision on to a broader societal impact, we are asked to face the scenario of a global catastrophic risk arising from the evolution of artificial intelligence. It is proposed that as no single way can be found, including alignment of human and AI, in covering all risks of AI, it may be more realistic to focus on the risk of narrow AI viruses and early self-improving AI in the sense that these 'risks are nearer in time and not overshadowed by other potential risks'. Continuing the ideological debate on global risk, we are introduced to the debate on the anthropocene as the 'contemporary counterpart of the Cold War doctrine of mutually assured destruction and the most compelling argument for a new kind of technological "arms race".' The discussion notes that in the era of an emerging ideological discourse on "energy security", the Anthropocene has come to represent the co-option of a 'scientific factography for the thinly disguised resurgence of "ideological science" of the Fukuyamaesque variety (posthistory, posthuman). According to this narrative, the argument goes, 'science—like technology—must be uniquely at the service of the maintenance of the global order, organised around a universal appeal to "crisis management".' It is this narrative that drives the posthuman trajectory towards an 'ever-more-apocalyptic Humanism'. Complementing the debate on the Anthropocene, it is argued that seen through the lens of ethical warfare in the form of ethical Lethal Autonomous Weapons (LAWs), artificially intelligent weapons systems can be so designed as to make decisions within the bounds of their ethics-based codes. The debate on such robotic weapons systems then posits that 'only ethical LAWs should be used to replace human involvement in war, and, by extension of their consistent abilities, should remove humans from war until a more formidable discovery is made in conducting ethical warfare', recognising that this philosophical argument may stimulate the building of a 'pragmatic strategy to 'ameliorate unnecessary violence'.

The challenge is thus to create a strategic framework that facilitates imaginative and creative response to technologies of the artificial, for example in dealing with the disruption of social, economic and cultural life, especially when life becomes synchronised with the digital environment. AI&Society welcomes contributions to the ongoing debate on Arts, Science and Society in our journal.

References

Cooley MJ (2019) The search for alternatives: liberating human imagination: a mike cooley reader. Spokesman, Nottingham **(forthcoming)**

Gill KS (2018) Artifcial intelligence: looking though the Pygmalion Lens. AI SOC Springer 33(6):459–465. https://doi.org/10.1007/s00146-018-0866-0

Weizenbaum J (1976) Computer power and human reason: from judgment to calculation. W. H. Freeman, Francisco

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