AI & Soc (2017) 32:475–482
DOI 10.1007/s00146-017-0755-y

EDITORIAL

Uncommon voices of AI

Karamjit S. Gill[1]

Published online: 14 August 2017

Beyond the headlines of the thrill engendered by futuristic AI super machines, Virtual Reality and Internet of Things, what are we to make of artificial intelligence? A gigantic job eliminator? Or the next step in evolution, the one in which technology finally asserts its mastery over us? Or maybe artificial intelligence in its many guises become the source of redemptive systems that develop new medications for us and operate on us, that invest and multiply our capital, and that create more rational decision-makers? (Ars Electronica Festival 2017). The new wave of artificial super intelligence raises a number of serious societal concerns: what are the crises and shocks of the AI machine that will trigger fundamental change and how should we cope with the resulting transformation? Digital technologies are the box in which we all increasingly live. Living through dramatic technological change, we may feel trapped and disrupted, being left behind in the myth and reality of AI, and miss what is really at stake. The Silicon Valley technological culture may often see societal concerns and humanistic perspectives of digital technologies as rather inconvenient, but in the midst of this transformation we can hear voices of existential risk, reason, redemption and ethics. Sir Rees (2013) of the Centre for the Study of Existential Risk (CSER) (2017) gives an insight into the concerns and challenges of existential risk of ecological shocks, fast-spreading pandemics, and scarcity of resources, aggravated by climate change. For him, equally worrying are the imponderable downsides of powerful new cyber-, bio-, nanotechnologies, and synthetic biology. His concerns include a "sci-fi scenario", in which a network of computers could develop a mind of its own and threaten us all. It is hard to quantify the potential "existential" threats from (for instance) bio- or cyber-technology, from artificial intelligence, or from runaway climatic catastrophes. He proposes forward planning and research to avoid unexpected catastrophic consequences and the imponderable downsides of powerful new cyber-, bio- and nanotechnologies, and to circumvent societal breakdown due to error or terror. Ó Éigeartaigh (2017) gives a soothingly rational note when he says that humanity has already changed a lot over its lifetime as a species. While our biology is not drastically different from what it was a millennium ago, the capabilities enabled by our scientific, technological, and sociocultural achievements have changed what it is to be human. We have dramatically augmented our biological abilities, we can store and access more information than our brains can hold, and collectively solve problems that we could not do individually. AI systems of the future would be capable of matching or surpassing human intellectual abilities across a broad range of domains and challenges. The Leverhulme Centre for the Future of Intelligence (CFI) (2017) visualises a redemptive curve on the horizon while asking us to take note of the serious consequences of untamed AI and argues for developing a framework for responsible innovation that seeks maximising the societal benefit of AI. He cautions us about the possibility of creating computer intelligence equaling that of human intelligence. In this future scenario, freed of biological constraints, such as limited memory and slow biochemical processing speeds, machines may eventually become more intelligent than we are—with profound implications for us all. Any inter-disciplinary or cross-disciplinary collaborative effort to meet these challenges,

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

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

he says, requires 'Value Alignment' for designing AI systems that do not inadvertently act in ways inimical to human values. As AI systems will operate with increasing autonomy and capability in complex domains in the real world, how can we ensure that they have the right behavioural dispositions—the goals or 'values' needed to ensure that things turn out well, from a human point of view?

In the very cognitively rational tradition of the Californian Silicon Valley, the Stanford Panel Report (2016) surmises that the frontier of AI has moved far ahead from the functions of the calculator, as AI researchers work on improving, generalising, and scaling up the intelligence currently found on smartphones. The report recognises that "Intelligence" remains a complex phenomenon whose varied aspects have attracted the attention of several different fields of study, including psychology, economics, neuroscience, biology, engineering, statistics, and linguistics. Naturally, it says, the field of AI has benefitted from the progress made by all of these allied fields. For example, the artificial neural network, which has been at the heart of several AI-based solutions, was originally inspired by thoughts about the flow of information in biological neurons.

The voices of rational reason keep reminding us that while new technologies of artificial general intelligence (AGI), synthetic biology, geo-engineering, distributed manufacturing will bring very large benefits to humankind, these also pose existential risks for human societies. Knight (2015) reminds us that the rapid developments in promoting machine learning and artificial neural networks modelled on biological networks have led to the debate on the existential threat posed by the future AI. He argues for the need to undertake proactive policy measures and a regulatory framework to mitigate the risks, even if no such breakthroughs currently appear imminent. Bostrom (2016) expounds that self-improving artificial intelligences could effortlessly enslave or destroy Homo sapiens if they so wished. While he expresses scepticism that such machines can be controlled, Bostrom claims that if we can program the right ''human-friendly'' values into them, they will continue to uphold these virtues, no matter how powerful the machines become. However, even if we recognise the limit of the super-intelligence machine, AI machines might still be extremely dangerous due to their potential for amplifying human stupidity. We are reminded that catastrophic threats are not merely academic—they actually do threaten humanity, and so for the sake of humanity they should be confronted. Baum and Tonn (2015) note that just as seeking generalised computational solutions to problems of existential risk may be tempting for machine learning ideologues, so is the idea of humanity living in simulations a computational fancy. They caution us on the danger of favouring safe AI technologies over dangerous ones, arguing that the standard ethical argument for confronting catastrophic threats to humanity is based on the far-future benefits of confronting the threats, rather than focusing on ''near-future" benefits from confronting near-future threats.

Amongst the conciliatory voices is that of Joi Ito, Director of the MIT Media Lab (2016), who cautions us about the exuberance of "extended intelligence," or E.I, as the dominant focus of AI on machine learning. Although AI scientists may be well intentioned in their building of machine intelligence tools, he says that "If we allow 'extended intelligence' to develop without thoughtfully managing how it integrates with, and affects, society, it could be used to amplify dangerous biases and entities". Unless AI scientists embed ethical and moral grounding in technology design and evaluation, the same technology that is meant to advance the well-being of society 'could, in fact, end up amplifying the worst aspects of our society.' For example, machine learning algorithms, under the guise of "smart machines", could be used to monitor citizens—to predict and project who would be a future criminal or a terrorist. Whilst the Internet has facilitated the springing of many social network movements, it has also increasingly become a place and platform for bigotry, hatred, prejudice, racism and malicious trolling. He agues for building technologies that, whilst being "smart", are also socially responsible. For this to happen, he argues for the development of "a framework for how our ethics, government, educational system and media evolve in the age of machine intelligence" by initiating "a broader, in-depth discussion about how society will co-evolve with this technology…". This voice is complemented by Jonathan Zittrain, co-founder of the Berkman Klein Center, when he says that "The thread running through these otherwise disparate phenomena is a shift of reasoning and judgment away from people, … Sometimes that's good, as it can free us up for other pursuits and and sometimes it's profoundly worrisome, as it decouples big decisions from human understanding and accountability for deeper undertakings. A lot of our work in this area will be to identify and cultivate technologies and practices that promote human autonomy and dignity rather than diminish it". (http://news.mit.edu/2017/mit-media-lab-to-participate-in-ai-ethics-and-governance-initiative-0110).

The voices of rational reason (Stanford Panel Report op.cit.) envision a future of developing systems that are human-aware, and focus on finding new, creative, interactive and scalable ways to teach robots, and bringing to bear the potential of AI and IoT-type systems for social and economic dimensions. The Stanford Panel predicts that in the coming years, new perception/object recognition capabilities and robotic platforms that are human-safe will grow, as will data-driven products and their markets.

These voices of rational reason would also have us see AI in terms of cognitive intelligence, arguing that 'the characterization of intelligence as a spectrum grants no special status to the human brain'. The argument is that any activity computers are able to perform and that people once performed should be counted as an instance of intelligence. As if AI were merely a technological wave of rational reason, the Stanford Panel found no cause for concern that AI is an imminent threat to humankind. The argument is that as no AI machines with self-sustaining long-term goals and intent have been developed so far, and are unlikely to be developed in the near future, we should focus on increasingly useful applications of AI, with potentially profound positive impacts on our society and economy in the near future. The Panel, however, recognises that many AI innovations would spur disruptions in how human labour is augmented or replaced by AI, creating new challenges for the economy and society more broadly. To mitigate the long-term impacts and consequence of AI, the Panel asks AI researchers, developers, social scientists, and policymakers 'to balance the imperative to innovate with mechanisms to ensure that AI's economic and social benefits are broadly shared across society'. It cautions the AI research community and policy makers not to take fears and suspicions of society lightly, and take steps to ensure the safety and reliability of AI. The Panel further asks them to engage society with a more open mind, if the technologies emerging from the field are to profoundly transform society for the better in the coming decades.

In the realm of voices of instrumental reason (https://cambridgeanalytica.org/), for data scientists, our brain is constantly required to adapt in a rapidly changing data-driven environment. When seen as predictive analytics, our brain is just a complicated learning machine whose main goal is data compression and interpretation. In this vision of data science, data processing, occurring automatically in our brains billion of times each second, is seen as an elementary step in many data analysis applications. Data science algorithms can be used to scan data for meaningful patterns, extracting combinations of features of meaningful data clusters. Beyond the voice of instrumental reason, Davies (2017) gives us an insight into the impact and implication of the shifting power of data, when he says that the majority of us are entirely oblivious to what all these data say about us, either individually or collectively. As personal data are becoming a huge driver of the digital economy, the data corporations are becoming 'more and more skillful at tracking our habits and subtly manipulating our behaviors'. In providing personal data to digital corporations in exchange for services, we are not only sacrificing our privacy rights, but in the process we are also allowing 'our feelings, identities and affiliations to be tracked and analysed with unprecedented speed'. He cites Cambridge Analytica (ibid.), which uses cutting-edge data analytics techniques, draws on various data sources to develop psychological profiles and targets millions of consumers with tailored messaging (e.g. targeting of American voters during the 2016 presidential elections). This ability to develop and refine psychological insights across large populations, he says, is one of the most innovative and controversial features of the new data analysis. He warns that in the world of data analytics where secrecy surrounding methods and sources of data is regarded as competitive advantage, it is doubtful that the 'big data elite' would easily give up their hold of data in favour of public interest and social benefit.

The voices of redemption point to the possibilities of mapping the landscape of potential AI breakthroughs and their social consequences. The argument is that keeping track of these developments will help to prioritise subsequent research, as control methods and social ramifications of AI will depend on both the system's architecture and the timeline for its arrival. From a boarder societal perspective, however, the future of AI poses challenges of democratic politics, including questions of political agency, accountability and representation. For example: with how well the existing institutions are equipped to deal with the risks and opportunities of the long-term transition to AI, and does AI require a technocratic rather than a democratic regulatory framework, and if so what might the cost be for democratic politics more widely (including for public confidence in democratic institutions)? This raises the question of how can machines be made politically answerable for their decisions in the way that human agents have traditionally been? If not, where is accountability to lie in any system when more and more of the work of government is being done by systems and machines? As artificial intelligence and robotics begin to fulfil their promise, they arrive pre-loaded with meaning, sparking associations—and media attention—disproportionate to their capacities. This matters: how we talk about new technologies and their risks and benefits can significantly influence their development, regulation and place in public opinion. Balancing AI's potential and its pitfalls, therefore, requires navigating this web of associations.

In their report, "When Will AI Exceed Human Performance? Evidence from AI Experts", Grace et al. (https://arxiv.org/pdf/1705.08807.pdf) of the Future of Humanity Institute, Oxford University note the massive social consequences of advances in artificial intelligence (AI). For example, self-driving technology might replace millions of driving jobs over the coming decade. In addition to possible unemployment, the transition will bring new challenges, such as rebuilding infrastructure, protecting vehicle cyber-security, and adapting laws and regulations. In

addition to the social and ethical impacts of AI, including the impact of AI and automation on human jobs, these transformative challenges, they say, will also arise from applications in law enforcement, military technology, and marketing. O'Reilly (2017) makes a plea for harmonising the new wave of AI for societal benefits. He asks us to take note of the voices of the rational proponents of AI, self-driving cars and on-demand services, and their commonality with 'income inequality'. He says that they are telling us, loud and clear, that we are in for massive changes in work, business, and the economy. We are heading "pell-mell" towards a world being shaped by technology in ways that we do not understand and have many reasons to fear. So what is the future, he asks, where is technology taking us? Is it going to fill us with astonishment or dismay? And most importantly, what is our role in deciding that future? How do we make choices today that will result in a world we want to live in? What is the future when more and more work can be done by intelligent machines instead of people, or only done by people in partnership with those machines? What happens to workers, and what happens to the companies that depend on their purchasing power? What is the future of business when technology-enabled networks and marketplaces are better at deploying talent than traditional companies? What is the future of education when on-demand learning outperforms traditional universities in keeping skills up to date? He further argues that we are at a very dangerous moment in history. The concentration of wealth and power in the hands of a global elite is eroding the power and sovereignty of nation-states at the same time as globe-spanning technology platforms are enabling algorithmic control of firms, institutions, and societies, shaping what billions of people see and understand and how the economic pie is divided. At the same time, income inequality and the pace of technology change are leading to a populist backlash featuring opposition to science, distrust of our governing institutions, and fear of the future, making it ever more difficult to solve the problems we have created.

We hear voices of ethics beyond regulation when Naughton (2017) alerts us to the social, ethical and legal implications of big data and machine learning. He cites the case of the transfer of health records of 1.6 million identifiable patients by The Royal Free hospital London to DeepMind, a Google-owned artificial intelligence firm, in July 2015, to create an app called Streams, to help clinicians manage acute kidney injury (AKI), a serious disease that is linked to 40,000 deaths a year in the UK (Powels and Hodson 2015). This collaboration in health care raised issues of on what ethical and legal bases did 1.6 million identifiable health records quietly disappear? And further, "How had the deal passed the various data-protection hurdles that any sharing of medical records have to surmount?" However, recently, when the UK Information Commissioner warned the Royal Free hospital on the non-compliance of the UK Data Protection Act, the DeepMind company conceded that they "underestimated the complexity of the NHS and of the rules around patient data…. We were almost exclusively focused on building tools that nurses and doctors wanted and thought of our work as technology for clinicians rather than something that needed to be accountable to and shaped by patients, the public and the NHS as a whole. We got that wrong and we need to do better" (Naughton 2017). Whatever future guidance the Information Commissioner comes up with, Naughton points out that "we are left with the fact that a database of 1.6 million sensitive health records that were transferred illegally is sitting on Google servers somewhere, even though DeepMind claims that it doesn't need it". What we take from this AI story is that we should be concerned about the myth that AI tools that affect the social fabric of society could be developed without abiding by the constraints of the legal, ethical, social, cultural values and norms of society. This example of DeepMind draws our attention to move beyond academic arguments on regulatory models when exploring the myths and realities of AI, of big data and machine learning, and promote, as an alternative, the creation of human-centred ethical frameworks that find a coherence between technological innovations and society.

As instrumental reason continues its march in the guise of machine learning algorithms, we see an increasing manipulation of data to support and control institutional and organisational structures. Moving beyond their (algorithms) role as computational artefacts, what concerns is how these algorithms take account of the limits of our 'entrenched assumptions about agency, transparency, and normativity'. Reflecting on these issues Gill (2017b) draws our attention to the work of observant authors, Introna, Crawford, and Ananny, who see data manipulation practices as problematic because they are inscrutable, automatic, and subsumed in the flow of daily practices. Beyond the issues of algorithmic transparency and openness, calculative practices have a serious impact on how domains of knowledge and expertise are produced, and how such domains of knowledge become internalised, affecting institutional governance. Moreover, these algorithms not only work within 'highly contested' online spaces of public discourse, they often perform with little visibility or accountability. This is an argument to move out of the 'black box' notion of the algorithm, and promote the idea of 'networked information algorithms' (NIAs); assemblages of institutionally situated code, practices, and norms with the power to create, sustain, and signify relationships among people and data through minimally

observable, semi-autonomous action. This opens the way for 'algorithmic ethics' that resembles 'actuarial ethics', based on the current and future risks. If AI reflections are to move out of the 'black box' of instrumental reason, we need to learn from the performance practices of artists, where performance of data is seen not just in terms of its transformation into information, but also in terms of the interactivity between the artist and the audience. This interactivity itself becomes a tool for the continued evolution of an artist and a scientist and the amalgamation of their partnership. In the end performance is about raising awareness of the interconnectivity of everything and everyone. Technology is or should be utilised to amplify the experience and/or the range of influence. As wearable sensors proliferate, we have access to rich information regarding human movement that gives us insights into our daily activities like never before. In a sensor-rich environment, it is desirable to build systems that are aware of human interactions by studying contextual information. Experiential scientists, crafts people, medical practitioners and engineers transform raw data into information, then using their skills and experience transform information into knowledge, and through the application of their contextual knowledge and wisdom, make judgements about the accuracy, relevance and acceptability of data that are coming from many sources. In this transformation process, there is always a scope for human intervention at various levels of the data-to-action cycle and that intervention, which reflects the many overlapping contexts, would bear witness to situated judgements. This is in contrast to an intervention based upon machine learning algorithmic calculations. In other words, the performance of data, in the hands of expert practitioners, is seen here in terms of an evolving judgement-making process culminating in action. This transformational process from data to action, encompassing feedback loops and human intervention, provides a human-centred perspective of judgement that is contrary to the computational model of 'judgement to calculation', in which data are used to compute judgement. We should, however, recognise that the computation model of judgement, turning judgement into an algorithm, is still a dominant focus of the data-driven AI. It may be tempting to argue that nothing has fundamentally changed in the data–action cycle except for the availability of an abundance of data (big data) and the exponential processing speed of computers. The fallacy of this argument then revolves around the idea that only if we have an abundance of data and the exponential processing speed of the computer, can we construct machine learning algorithms that can outstrip human cognition, to the extent that machines can better humans in processing a wider variety and larger number of data sets and working in different ways to those of humans in reaching analytical judgements. However, this calculation-centred view of judgement fails to recognise that human judgement is about the process of finding a coherence among often conflicting and yet creative possibilities that cannot be reduced to calculation. Moreover, human judgement resides in and reflects the dynamic and evolving nature of professional and social practices, enriching human experience, knowledge, skill and cognition. From this human-centred perspective, performance of data lies in the performance of practice of the 'data–action cycle', in other words the performance of inter-relations between data, information, knowledge, wisdom and action (Gill 2017b). This view seeks to understand the nature of the interface between the physical, cultural and our experiential worlds. The nature and practice of the interface here is fundamentally relational between, in-between, and across knowledges, experiences and practices of contextual domains (Gill 2015), and not transactional in the sense of 'cause and effect' calculation. This view shifts our attention from a purely technological fascination of machine learning to the evolving interaction of human systems and technology, thereby providing a symbiotic horizon of performing data. In the midst of the fascination with digital technology, we are cautioned to remember that performance of data in the hands of creative artists and scientists embodies social/cultural and spatial intelligence that conforms to the living. We cannot get this from machine intelligence. Moreover, it is not clear how a machine would deal with the architectural paradox: when an architect draws a diagram of a building, the diagram becomes a building, a static object, an exact language, an exact dream; but the diagram as a model performs as a process, a dynamic process in which the diagram acts an algorithm of ideas. Such a discussion on the creation of an ethical framework needs 'to be infused with a more robust notion of the public interest than can currently be found in the realm of digital intermediary governance' (Gill op.cit).

AI & Society authors in this volume add to these uncommon voices of AI from their own perspectives, thereby contributing to the ongoing exploration of socially responsive developments of AI. Among the voices of the uncommon, Danila Bertasio (this volume) argues that man continues to play the imitation game, fantasising replication of the self, blurring the line between the natural and the artificial, even at the cost of breaking cultural boundaries and taboos. However, this dream of true replication seems to exhibit signs of disillusionment and subsequent abandonment in meeting the standards of contemporary technical advancements in robotics. The modern-day engineer pursues the same dream as did the fourteenth-

century wax-workers, unaware that doing so is related to a kind of neo-Platonism that might be even less tolerant of a copy—namely, to construct simulacra that would be expected to behave 'humanly'. Overcoming the disquieting effect of the wax-workers' statues, robotics once again pursues the dream—in constructing robots that would ideally be indistinguishable from humans—of building a replicant very similar to its creator. Man's attempt to create a replica of himself, through the unification of technological and aesthetic levels of observation, has deep roots dating back at the very least to the artifices of Heron of Alexandria and his teacher Philo; and today's anthropomorphic robotics shares the same replicative philosophy, albeit with an interesting underlying difference. Indeed, while the ancient automata had mainly recreational, imaginary or mythical purposes, as did those of the eighteenth and nineteenth centuries, today's anthropomorphic robot design seems to be aimed at creating a 'perfect' double—an endeavour that proposes a curious continuity with others presenting themselves throughout history, such as, for example, the production of wax 'doubles' in the fourteenth century. However, the fate of such desires seems ineluctably sealed. In the background there is always the same constitutive limit that characterises the human condition that consists in the obstinate tendency to replicate without an accurate knowledge of the object to reproduce.

Could the robot, as the replication of man, reach technological sophistication such that the robot could "nudge" a user's behaviour for the good of society? Jason Borenstein (this volume) explores the creation of companion robots that would seek to nurture a user's empathy towards other human beings. Could a companion robot encourage humans to perform charitable acts, and could it potentially elicit from a user what the associated ethical concerns may be? This nudging behaviour for social good raises a number of questions, for example: who determines what is good for society in this context? Are there any universal social goods that should be considered? What role, if any, do cultural variations and tolerances have in this context?

In the pursuit of social good, Sofia Serholt et al. (this volume) explore the potential of the robot to facilitate children's learning and to function autonomously within real classrooms in the near future. In response to ethical concerns surrounding children interacting with robots, the authors draw on a Responsible Research and Innovation perspective, and discuss the design of features that will render robots more socially acceptable, taking account of teachers' perspectives on classroom robots pertaining to privacy, role of the robot, effects on children, and responsibility. It is suggested that beyond privacy, intentional or unintentional consequences, robots could potentially affect children in negative ways, whereby the risks are considered to outweigh the possible benefits. This raises the issue of who could be held accountable for negative consequences, and what responsibility do designers have in designing social robots?

What social challenges do developers have to face in promoting safe and beneficial artificial intelligence? Seth Baum (this volume) sheds some light on the motivation and measures for making a shift from the current focus on developing capable AI towards building more beneficial AI. Extrinsic measures impose constraints, incentives or compliance on AI researchers to induce them to pursue beneficial AI even if they do not want to. Intrinsic factors such as social contexts and social meaning, social norms, contextual messengers and allies encourage AI researchers to want to pursue beneficial AI. And framing can both determine the success of extrinsic measures and motivate AI communities to develop beneficial AI. Baum (this volume), however, alerts us to the dangers of extreme framing: framing of strong AI as a powerful winner-takes-all technology, makes a supposedly dangerous technology seem desirable; framing of AI researchers as people who do not want to pursue beneficial designs, can potentially be counterproductive; and extreme proposals like draconian global surveillance can inadvertently frame efforts to promote beneficial AI as being the problem, not the solution. In other words, it could give the impression that the efforts are misguided and causing more harm than good. Stigmatisation is a type of framing oriented towards making an object or an activity feel socially undesirable or even taboo. Stigmatization can be an effective technique for preventing the use of dangerous technologies, and can also be used for both rejecting claims of AI harm and for promoting beneficial AI. The aim of any measure should be to reduce the harms and increase the benefits of AI to society. A measure that does this should be pursued, even if it still leaves some potential for harm or for loss of benefit. Given the stakes involved in AI, all effective measures for promoting beneficial AI should be pursued.

Inserting some relief to the serious debates on existential risk of uncommon AI, Huma Shah and Kevin Warwick (this volume) look at the possibility of a machine having a sense of humour, contrary to Turing's 'arguments' from various disabilities' used against the concept of a machine being able to think. We can envision social robots performing a sense of humour while being able to be rude and sometimes even offensive, although this can depend on their interrogators and how sensitive they are.

Can social robots be conscious when performing humour or rudeness or can AI machines be conscious when performing social responsive acts or acting harmfully. Rajakishore Nath (this volume) asks whether unintelligent machines could give rise to an intelligent conscious experience, have the perception of thought, feel and have awareness. Nath argues that the causal explanation of the

'how' and 'what' of consciousness fails to explain the 'why' of consciousness. Situated in the mechanistic framework of the sciences, this epistemological theory of consciousness is essentially committed to a scientific world view that cannot avoid the metaphysical implication of consciousness. Towards this perspective of consciousness, Nath introduces us to the neo-*Advaitins* who have maintained that the evolution of nature leads to the manifestation of human consciousness, only because consciousness is already implicit in material nature. Thus, the existence of consciousness in this physical world far exceeds the methods of science and needs a non-mechanical metaphysical explanation.

Taking inspiration from the Batesonian ecology of mind, Floridi's information ethics, Felix Guattari's ecosophy, Braidotti's posthumanism, and the Japanese animist doctrine of Rinri, Vassilis Galanos (this volume) explores the nature of consciousness (the natural/artificial dichotomy) and the future of artificial agency as a potential existential threat. This exploration covers human–robot cultural contact, from the early scientific discourse of Man–Machine Symbiosis up to the contemporary counter-measures against superintelligent agents. Vassilis ponders on Bateson's double-bind theory acting as the "therapeutic double bind," to confronting messages of proponents and opponents of artificial intelligence and humanity's conscience of habitualizing danger and familiarisation with its possible future extinction. He surmises on the dilemma of getting caught up in a double bind whose therapy reveals the possibility of the patient species' gradual extinction, for the development of higher forms of intelligence (if applicable). It is like a meta-double bind, where one is caught in the middle of either staying in therapy to cope with one's possible extinction, or returning to the initial bind of the lose–lose scenario.

Mikael Wahlström (this volume) explores the role of public imagination for acceptance of future technologies of automated transportation. Authors suggest that public imagination, along with media discourses and societal settings that contribute to explanations, should be considered in the design and study of automated systems. Moreover, social representations could be beneficial for media frame studies by providing explications as to why certain frames might have or lack cultural resonance.

Douglas Walton and Marcin Koszowy (this volume) examine the problem of the uncritical acceptance of expert opinions that is at the root of the ad verecundiam fallacy, and argue for the need to disentangle argument based on expert opinion from another kind of appeal to authority. In dealing with this fallacy, they shed light on the argument from expert opinion as it concerns reasoning about how things are, as in theoretical reasoning, as well as in the other type of authority labelled 'deontic' or 'administrative'. They draw a distinction between the two types of authority, expert and deontic. In the case of epistemic or cognitive authority, the domain of authority is a set of propositions which are asserted, e.g. by an expert in a given field. In the case of deontic or administrative authority, the domain of authority consists of, e.g. commands, requests and advice. The capability to systematically distinguish between these two types of argument from authority has been shown to open up new avenues for investigating the more serious instances of the ad verecundiam fallacy where the two types of argument are systematically confused. The authors argue that formal and computational argumentation systems enable us to analyse the fault in which an error has occurred by virtue of a failure to meet one or more of the requirements of the argumentation scheme from argument of expert opinion. The essential characteristic of the sophistical tactic type of ad verecundiam fallacy consists in a sequence of moves in a dialogue fitting the pattern of a device to force premature closure of the dialogue. Ultimately, they conjecture, full analysis of the ad verecundiam fallacy will not be achieved until the dialectical properties of this kind of argumentation can be modelled.

Luo and John-Jules Meyer (this volume) examine the use of formal models to explore the notion of opportunistic behaviour in social interactions. This includes the way situation calculus can help in understanding this behaviour, and gaining insights into the compatibility of different value systems and the co-evolution of agents' value systems with social context or environmental changes. They consider opportunism that would cause harm to others and that for gaining personal advantage. Seeing opportunism as a self-interested behaviour that conflicts with social norms, they suggest that its emergence might come from the way in which agents resolve the conflicts between beliefs, obligations, intentions and desires. They propose that similar to lie-detection, a well-designed monitoring mechanism can be used to automatically detect opportunism in (computer-based) human interactions, thereby providing ways to protect agents' values from being demoted. Further, the monitoring mechanism could include the design of constraint mechanisms that eliminate or prevent opportunism from happening.

In the wake of technological voices for social good, Devendra K. Tayal (this volume) explores the application of sentiment analysis to shape social campaigns effectively for the betterment of the society.

Continuing the performance of technological voices for social–economic good, Gagan Deep Kaur (this volume) explores the way technological interventions have over the years shaped the technological makeover of the design of artefacts and triggered major changes in the practice. This has resulted in heralding profound cognitive accomplishments in the manually driven process of Kashmiri carpet

design, causing major alterations in the overall structure of the practice. She notes that the recent intervention of digital technology has, on the one hand, brought precision and speedy processing in the design process, and on the other hand, it has eliminated some of the crucial actors from the practice, thereby having a cognitive impact on the design process as well as on the practice.

In many ways these voices of AI are a continuation of the human-centred debate of the 1970s (Gill 1996) that is rooted in the idea that machines calculate and humans make judgements. Human-centred thinkers of the 1970s (Cooley 1987; Rosenbrock 1990) felt perturbed about the hold of the scientific method of Taylorism and its implications for working life in the industrialised world, and by implication for the wider society. The concerns included the fear of the automation of production processes, the mechanisation and by implication de-humanisation of the work place, the loss of human skill and expertise, and ultimately the replacement of the human worker by the robot, leading to mass unemployment and exclusion. At the same time horizon, the computer as a symbolic embodiment of instrumental reason was seen to go further than the machine, being made in the image of man, an imitation of a certain aspect of man in the sense that it ventured into the realm of the imitation of human thought (Weizenbaum 1976). This was seen as a step towards the reproduction of some key aspects of human traits if not their replacement. There was further unease at the idea of venerating the machine to the point that there is no difference between humans and machines, and between human thought and machine thought. Whilst voices of reason and rationality perceive the AI wave in the pursuit of reality–reality interactions, Uchiyama (2003) from a Japanese perspective sees this wave in the pursuit of reality–actuality relations. Uchiyama draws a distinction between the way Western and Japanese participants comprehend situations. While the Western participant 'sees' the situation and relates to it as an objective observer, the Japanese participant "hears" the situation, and relates to it by feeling to be "in the situation". In the first case, the interaction between the observer and the situation is through information, and in the second case the communication between the observer and the situation is through language. This transformation from information to communication provides a dialogical voice of reflection (Gill op.cit) on the AI wave. Various perspectives of the myths and reality of AI are explored in the international journal, *AI & Society* (Gill 2016a, b, 2017a, b).

AI & Society warmly welcomes reflective contributions to the debate on uncommon voices, exploring the myths and reality of AI in the pursuit of seeking harmonious interactivity of art, science, technology and society.

References

Ars (2017) The 2017 Ars Electronica Festival, Linz, Austria, September 7–11, 2017. The theme: artificial intelligence. https://www.aec.at/news/en/festival2017/

Baum SD, Tonn BE (2015) Confronting future catastrophic threats to humanity. Futures 72:1–3

Bostrom N (2016) Superintelligence: Paths, Dangers, Strategies. OUP Oxford, Oxford

Centre for the Study of Existential Risk (CSER) (2017). http://cser.org/. Accessed 28 July 2017

Cooley MJ (1987) Architect or Bee?. Hogarth Press, London

Davies W (2017) How statistics lost their power—and why we should fear what comes next. The Guardian. https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy. Accessed 28 Apr 2017

Gill KS (ed) (1996) Human machine symbiosis. Springer, London, p 1996

Gill SP (2015) Tacit Engagement: beyond interaction. Springer, Berlin

Gill KS (2016a) Artificial super intelligence: beyond rhetoric. AI Soc 31:2. doi:10.1007/s00146-016-0651-x

Gill KS (2016b) Performing ethics. AI Soc 32:1. doi:10.1007/s00146-016-0687-y

Gill KS (2017a) Preface—hermeneutic of performing knowledge. AI Soc 32:2

Gill KS (2017b) Hermeneutic of performing data. AI Soc 32:3. doi:10.1007/s00146-017-0727-2

Grace K et al. When will AI exceed human performance? Evidence from AI experts. Future of Humanity Institute, Oxford University. https://arxiv.org/pdf/1705.08807.pdf. Accessed 23 May 2017

Ito J (2016) Well-intentioned uses of technology can go wrong. MIT Media Lab. https://www.nytimes.com/roomfordebate/2016/12/05/is-artificial-intelligence-taking-over-our-lives/well-intentioned-uses-of-technology-can-go-wrong

Knight W (2015) What robots and AI learned in 2015. MIT technical review, December 29, 2015 http://www.technologyreview.com/news/544901/what-robots-and-ai-learned-in-2015/. Accessed 5 Jan 2016

Leverhulme Centre for the Future of Intelligence (2017). http://lcfi.ac.uk/. Accessed 13 July 2017

Naughton J (2017) Giving Google our private NHS data is simply illegal. The Guardian, Sunday 9 July 2017. https://www.theguardian.com/commentisfree/2017/jul/09/giving-google-private-nhs-data-is-simply-illegal

Ó Éigeartaigh S (2017).Technological Wild cards: existential risk and a changing humanity. https://duckduckgo.com/?q=Technological+Wild+Cards%3A+Existential+Risk+and+a+Changing+Humanity&t=ffhp&ia=web

O'Reilly T (2017) The WTF Economy. CRASSH, University of Cambridge, 23 May 2017

Powels J, Hodson H (2015) Google DeepMind and healthcare in an age of algorithms. Health Technol. Springer. doi:10.1007/s12553-017-0179-1. https://link.springer.com/content/pdf/10.1007%2Fs12553-017-0179-1.pdf

Rees M (2013) Denial of catastrophic risks. Science 339(6124):1123. doi:10.1126/science.1236756

Rosenbrock H (1990) Machines with a purpose. Oxford University Press, Oxford

Stanford University (2016) Artificial intelligence and life in 2030. One hundred year study on artificial intelligence: report of the 2015–2016 Study Panel, Stanford University, Stanford, September 2016. http://ai100.stanford.edu/2016-report. Accessed 9 July 2017

Uchiyama K (2003) The theory and practice of actuality. Institute of Business research, Daito Bunka University, Tokyo, Japan

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