AI & Soc (2016) 31:137–143
DOI 10.1007/s00146-016-0651-x

EDITORIAL

# Artificial super intelligence: beyond rhetoric

**Karamjit S. Gill**[1]

Published online: 11 February 2016


> Logic will get you from A to B. Imagination will take you everywhere
> -Albert Einstein

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? What would the implication be if AI machine takes over and transforms the way we live and work? What would technology do to work, employment, economy, governance, state, democracy and professions? What would the social and political implications of employment be if people were replaced by the machine? What if the state disappears, and so do economy, professions, employment, politics disappear with it as we know them? Can digital economy be regulated, measured, and controlled? Can the AI machine with its embedded machine learning algorithms be monitored and controlled? Would new politics emerge as another digital game, and what would the rules of this game be, and how would these rules change the playing field of the game of politics itself? Would the nasty form of exploitative individualism triumph or would new forms of digital collectives (e.g. consumer collectives) emerge that would be more powerful than corporations? And can humanity live in a simulated state of digital being? These and many such questions arose during a recent symposium on "Technological Displacement of White-collar Employment: Political and Social Implications" held at Cambridge University (CRASSH 2016).

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

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

These questions raise further issues: Which of the things lie beyond the capacity of the machine or should a human be expected to accommodate the machine, for example conforming to the digital path of the Google autonomous car. What tasks should not be handed to and handled by the machine? For example should a machine turn off the life support system? Again how can we understand the reasoning of the AI machine and how do we interpret this reasoning process, given that the machine thinks differently from a human being? Is it a mistaken assumption or a belief in the bounded rationality or both that the logical and probabilistic rule-based reasoning AI machine can think and act as humans do? Or are we all expected to be seduced to fit into the autonomous good path of the AI machine?

It is perhaps not surprising that such a questioning of the super-intelligent machine attracts rather a tired response from even well-meaning researchers, which goes along the lines: Give me a break, don't bother me with social responsibility sermons! I am only a researcher interested in the creative and disruptive innovation. I build tools and systems, it is up to society whether to use them for good or evil. Technology is neutral, it is humans who are not. At the same time, it is heartening to note that there are many researchers and policy makers who argue for an ongoing conversation between humanity and technology in order to reflect and plan what could be and what should be done in the time of crisis triggered by the autonomous AI machine? And what would this conversation look like? Should we strive to create alternative digital collectives to shape the digital transformation? And what would be the social utility of these collectives, for example in caring, welfare and recreational economies?

The questions and issues cited above are seen as part of the wider debate on AI and existential risk, autonomous robots, Big Data and the Internet of Things. We are minded that while new technologies of artificial general

intelligence (AGI), synthetic biology, geo-engineering, distributed manufacturing will have very large benefits to humankind, these also pose existential risks for human societies. Knight (2015) says that 2015 was the year of debates on self-driving cars, robotics, deep learning, and super-artificial intelligence. These rapid developments, promoting machine learning and artificial neural networks modelled on biological networks led to the debate on existential threat posed by the future AI. These risks include "the creation of new weapons of mass destruction, or catastrophe through accidental misuse". Moreover, artificial general intelligence "underlies human capabilities in strategising, social manipulation, hacking, technology research, and economic productivity". Since the nature of this technological progress is unpredictable, there is a need to undertake proactive policy measures and a regulatory framework to mitigate the risks, even if no such breakthroughs currently appear imminent. It is noted that Bostrom's study of "existential risk" (Future of Humanity Institute 2013, Knight ibid.) argues that artificial intelligence might be the most apocalyptic technology of all. With intellectual powers beyond human comprehension, Bostrom 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 program the right "human-friendly" values into them, they will continue to uphold these virtues, no matter how powerful the machines become. Commenting on rhetoric of deep learning, Knight further notes that at the core of deep learning lies the limit of artificial neural networks, in the sense that these artificial neural networks cannot compute at the speed or accuracy that our brains do. Thus one of machine learning's most intractable limits is the development of artificial neurons that can function at an accelerated rate.

Commenting on Bostrom's study on existential risk, Geist (2015) says that while recognising the limit of the super-intelligence machine, AI-enhanced technologies might still be extremely dangerous due to their potential for amplifying human stupidity. So far as the existential risk is concerned, just by enhancing the familiar twentieth century technologies, AIs of the future can endanger the future survival of existing societal structures by undermining their precarious strategic balances, for example by making the existing technologies much faster, cheaper and deadlier. If anything, Geist says that machines capable of conceiving and actualising elaborate plans but lacking self-awareness could be far more dangerous than mechanical analogues of human minds.

Baum and Tonn (2015) note that the bulk of the catastrophic threats literature has thus far focused mainly on philosophical aspects, in particular the moral significance of catastrophic threats and challenges to quantifying their probability, as well as empirical aspects, in particular the nature and size of the various threats. Although there is considerable research into specific threats such as global warming and nuclear war, there is rather a lack of much needed research into existential risk. They further note that catastrophic threats are not merely academic—they actually do threaten humanity, and so for the sake of humanity they should be confronted. For example, there is a need for the "better development of Quantum-safe encryption and its wider deployment to avoid spying on citizens, corporations, and countries, potentially enabling catastrophic totalitarianism and economic chaos".

On the threat from advanced artificial intelligence, Baum et al. (ibid.) note the potential for global government ("singleton") and for the possibility that humanity exists within a computer simulation. 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. We need to be mindful of the differential technological developments in which safe AI technologies are favoured over dangerous ones. Baum et al. present a practical perspective on the ethics of catastrophic risk. They articulate that the standard ethical argument for confronting catastrophic threats to humanity is based on the far-future benefits of confronting the threats. They posit that those who do not yet share the argument on existential risk may contribute to long-term research by focusing on "near-future benefits from confronting near-future threats, as well as "mainstreaming" actions on the threats into existing activities. They survey the threats, finding that "probably a large majority" of the total threat can be confronted with actions that appeal to "what people already care about", and furthermore that these actions "will often be the best to promote, achieving the largest GCR (Global Catastrophic Risk) reduction relative to effort spent".

On the potential and implication of big data, Naughton (2015) in a recent opinion column raises issues of privacy and security of the technological juggernaut of the Internet of Things (IoT). He notes that for the tech industry, it is the Next Big Thing, alongside big data, though in fact pair are often just two sides of the same coin. He says that the basic idea is that since computing devices are getting smaller and cheaper, and wireless network technology is becoming ubiquitous, it will soon be feasible to have trillions of tiny, networked computers embedded in everything. These can sense changes, turning things on and off, making decisions about whether to open a door or close a valve or order fresh supplies of milk, you name it, the computers will be communicating with one another and shipping data to server farms all over the place.

Central to the design of AGI and Big Data systems is the challenge of what is it to be human in the face of these

emergent technologies? This goes to the heart of human–machine relations, the place of technology in human systems, and the human place in technological innovations. The debates on AI and Big Data bring to focus some of the key issues, which have a bearing on the design of super-intelligent systems. For example, the great potential of AI technology and Big data is accompanied by a professional, ethical and moral responsibility to protect the safety of those around us. The Responsible Data Forum (2015) emphasises that responsible data are not just about technical security and encryption, but about prioritising dignity, respect and privacy of the people we work with, and making sure that the people reflected in the data we use are counted and heard, and able to make informed decisions about their lives. A concern is that just as users face algorithmic biases, and especially when algorithmic solutions are promoted for human domains including social risks and disasters, regulatory frameworks designed for global solutions tend to exclude personal, local and community contexts and even regional and national contexts, thereby leading to the exclusion of the very people who are at risk. The challenge, according to this Forum, then is how to manage the risks and achieve great successes and gains, while still being sensitive to asymmetries. In addition to the promotion of generalised rules, there is also a concern about the increasing reliance on quantification: documenting, measuring, monitoring and reporting, at the cost of the genuine rights and concerns of people. Moreover, "Future-proofing" is also an issue: "What seems unproblematic data right now, for example, may turn out to be very sensitive in the future".

Boyd's (2015) view on the "buzz around artificial intelligence" is that "the central issue of the twenty first century is not machines taking over, it's how to achieve the right balance between humans and automation to optimize outcomes". He makes a point that the most valuable resource we have is human intelligence, in other worlds human skill, ingenuity and creativity. On the other hand, machine intelligence is information and computation of data. It is through a symbiotic balance between human intelligence and machine intelligence that we can achieve optimal solutions to problems, which matter most to humanity. Taking the example of jets replacing propeller-driven aircraft, Boyd illustrates that "the key to successfully flying a jet aircraft was learning what to outsource to the automated systems and what to retain for human management". Again, he says that in the case of F35 aircraft, the pilot's helmet links the pilot with the sensors around the plane that allow the pilot to actually be able to "see through" the aircraft, thereby describing the aircraft as an extension of themselves and in cases desiring the experience as if they were fused with machine, a true symbiotic relationship between the human and the machine. It is through this fusion, a "fluid interface" of symbiotic collaboration between the human and the machine, Boyd says, that we can build new capacities that the human alone or machine alone would not be able to accomplish.

Contributions to this volume subscribe to some of the issues raised above, ranging from potential and risks of AGI, autonomous robots, simulation of human civilisation, and auto-catastrophe's manifestation. Discussions in this volume bring to the fore some of the key topics which should be of interest to designers of intelligent interactive tools and systems. These range from enactive appropriation, interpretation, simulated existence, interface paradigm, conversational judgement, evolutionary telos, culture and technology transfer, and adaptive learning. What is common to these topics is symbiotic interface between technology and society, a relational symbiotics, emphasising the contextual use and affordances of tools and systems.

Phil Turner in his insightful discussion on, "Presence: Is it just pretending?", makes us ponder on the relational significance of presence in human engagement and thereby on the design of affective computer mediated interfaces. He makes us reflect on our presence in the real social world or our make-believe presence in other worlds or virtual worlds. The experience of presence or immersion in a movie, game or virtual environment is not automatic but is the product of our deliberate engagement with it, an engagement which first requires a disengagement or decoupling with the real world. The argument is that the means by which we feel present in these other "worlds" lies in our ability to make-believe. These make-believe or imagined alternate worlds may be not as vivid, immediate, or as tangible as the real world, but they can be very engaging, especially when mediated through external artefacts such as toys, books and works of art or more recently with digital technology. Turner says that making-believe is a form of cognition which is decoupled from the real world and which enables us to explore and engage with fictional or imaginary worlds. If make-believe opens the door to other worlds, then the sense of mediated presence keeps it open. Turner further notes that the power of make-believe is not to be underestimated, as it is astonishingly ubiquitous and can be found at work in everything from the kind of mundane "what if" thinking we might engage in when deciding what to have for dinner, through to scientific reasoning (e.g. Einstein famously imagined himself chasing a light beam). Taking a cue from Turner, we should be mindful of how our sense of presence in the real world helps regulate our behaviour within it, and how this ability offers us practical advantages in dealing effectively with the world.

These reflections on presence with mediating artefacts may be seen as the continuation of, or complementary to,

the discussion on 'Enactive Appropriation' (Flint and Turner 2016), which goes to the very heart of interface design as being beyond the appropriation of artefacts, to being about the use which transcends the original design of the artefact. The authors argue that it is appropriation through the manipulation of affordances of tools that enables us to seek user relationships with digital technologies in a way that the potential of this knowledge can positively affect more functional applications. Use can thus be seen as making a creative, effective and purposive appropriation of tools, systems and resources beyond their original conception and design. In other words, appropriation emerges as a natural consequence of this enactive use. Perception, from these perspectives of Enaction, is an active process. It is something we do, and not something that happens to us. From this reading, authors posit that, use then becomes the active exploitation of the affordances offered us by the artefact, system or service. In turn, authors define appropriation as the engagement with these actively disclosed affordances—disclosed as a consequence of, not just, seeing but of seeing as. This enactive approach treats perception as an active skill, meaning that perception involves active exploration of the world rather than "interpreting the patterns of light falling on the eyes". The authors suggest that enactive (visual) perception is not so much a matter of seeing but of seeing as. They further note that we perceive the affordances offered by the world and act on those, or as Heidegger has put it, we encounter the world as ready-to-hand and available. This account of appropriation stands in sharp contrast to theoretical treatments because it rejects a role for representation, schemata or mental models.

Silva's discussion on "Human, Machines, and the Interpretation of Formal Systems" emphasises the ever presence of the intelligent machine in the form of digital and autonomous robots. It brings to our attention the centrality of the interpretation of automatic formal systems (programs running on a digital computer), which lies at the heart of these intelligent machines. The argument is that if the interpretation of a formal system does not belong to the formal system itself and if the interpretation has to be added, the question arises as to where does the interpretation come from? It notes that the human source of interpretation of formal interactive systems is sometimes concealed by a formalist restriction. Commenting on this invisibility of interpretation, the discussion underlines our responsibility, as human agents, for all this interpretative work—and its importance for us as human beings. The automatic formal systems controlling autonomous robots are more grounded in our world than automatic formal systems controlling desktop computers, in the sense that robots have sensors and motors making them able to sense and impact external world beyond symbols. However, when it comes to the meaning of social robots (e.g. taking care of an old person), it depends not only on sensors and robot actuators; it also depends on the continued sharing of a group life, which is (for now) out of the reach of these machines. The author argues that we can realise better the interrelationship of social robots to our social world, if we understand that the connection between a formal system and the world is given by the interpretation of the formal system, and this should be our responsibility as designers.

White's discussion on "Simulation, self-extinction, and philosophy in the service of human" summarises the notion of "ancestor simulation" in the realm of the super intelligent machine—that is as a simulation of a period prior to that in which a civilisation more advanced than our own—"post-human"—becomes able to simulate such a state of affairs as ours—"human". It sheds light on ancestor simulation by exploring the motivating rationale behind current work in the development of psychologically realistic social simulations. By rendering human cognition in a computational medium, dynamic system models of cognitive agency can reproduce many aspects of human systems that may in other forms be considered incomputable, i.e. political voice, predictive planning and consciousness. In this sense, the author argues that simulations afford a unique potential to secure a post-human future that may be necessary for a pre–post-human civilisation like our own to achieve and to maintain a post-human situation. The discussion posits that worries that had prompted research into AI and existential risk, that we may inhabit a simulation for the entertainment of evil demons, can be laid to rest. The world is just as it must be, and we must do something about it. Rather than continue to live, predict, plan and act as normal, we should accept it as our constitutive purpose to correct error. The trouble is simply that we are the only ones present to fulfil this purpose. The future—the only real future facing us unless living under constant threat of self-annihilation can be counted as a "future" at all—exists in adapting technologies in the optimisation of social and political systems for problem-solving, sustainability and ultimately the good life. To this end, the future of philosophy, the author asserts, arises at least in part in the grounding of the cognitive and the social sciences in the physical sciences, in understanding the metaphysical in terms friendly to realistic simulation, at the very least so that emerging simulations can be measured against an enlightened and guiding account of the human condition. Through this industry, stable visions of possible futures and the paths that take us there may be discerned, proactive human self-direction past extinction-level threats may be facilitated, and post-human political potential may be realised. In the end, the probability of a simulated existence cannot be grounded in the possibility of a post-human existence accidentally interested in an ancestor simulation

for which we have no evidence beyond speculation. Rather, the probability is proportional to the clearest and most distinct direct evidence imaginable, our own felt commitment to similar ends through similar means, today. Given that we are not the first in the universe to find ourselves in such a situation, and as a simulating civilisation should create countless more simulations than exist as actual worlds, the sum of our collective commitment is proportional to the probability that we live now in a simulation set up in a similar effort by someone else perhaps a very long time ago.

Kyriakidou's discussion on "Auto-Catastrophic Theory" provides an evolutionary perspective of existential risk, arguing that our survival would have been impossible without the presence of self-destruction in multiple levels of our existence and surroundings. The argument is that without the existence of the auto-catastrophe of living systems that exists within cells, people, societies, earth and galaxies, humanity may not have existed, or may not have been able to function, or an organism may not have been able to live. Self-destruction, the author says, goes beyond the development of a system; it contributes to the development of entirely new systems. This makes auto-catastrophe both a positive and a negative mechanism. However, self-destruction is highly related to the needs of the system to survive and to the need of the system to end. So how do we envision the place of "non-alive" systems such as artificially intelligent devices, and nanotechnology, and can these "non-alive" systems be used to save evolutionary telos of humanity? And can humanity be defended from such deuterogenic (e.g. life extension) auto-catastrophic processes by programming protogenic (e.g. death) auto-catastrophic processes into them. The assertion is that non-"alive" systems (e.g. nanotechnology) can be used to increase deuterogenic survival processes for people, and to turn people into partially non-"alive" systems to overwhelm protogenic auto-catastrophic processes (e.g. death). Non-"alive" systems can be a threat to humanity when they are considered to be partially "alive" systems (e.g. artificial intelligence) and they can also conduct protogenic survival processes, because they are not vulnerable to protogenic auto-catastrophic processes but only to deuterogenic auto-catastrophic processes. The more we turn to artificial intelligence to help our survival the greater is the threat (by artificial intelligence) of our extinction. This is because, first, artificially established devices will develop protogenic survival processes turning them into a potential threat for us (as a deuterogenic auto-catastrophic process). Second, the more we try to overcome protogenic auto-catastrophic processes (e.g. death) the less "humans" we become (so we can achieve partially "alive" systems that are not vulnerable to auto-catastrophe).

Armstrong et al.'s discussion on "Racing to the precipice" gives an insight into the possible competitive nature of AI teams working on existential risk, in the sense that such a competitive research culture can increase the danger of an AI-disaster, especially if risk taking is more important than skill in developing the AI. It posits that information can also increase the risks: the more teams know about each others' capabilities (and about their own), the more the danger increases. By illustrating the example of competitive arms race, it points to methods of increasing the chance of safe development of AI, and reducing the existential risks. It, however, notes that in relation to the arms race, AI risk seems to be in a unique category. The argument is that game theoretical analyses of other disasters, such as climate change, focus on the unwillingness of participants to take on a personal cost for a general good, where there is broad value alignment but divergence of information. In contrast, the authors say that AI potentially poses great risks to its creator, whereas nuclear powers are not generally at direct risk from their own arsenals. They note that may be the closest equivalent is the potential risk of accidental release of pathogens in biotechnology, where the rush to be first with some major discovery could cause the teams to skimp on safety precautions. In a less apocalyptic vein, it could be applied to many scenarios where there is pressure to achieve a "first" and the potential for loss if this is not done carefully (certain medical or food related products could fall into this category).

Warwick and Shah's discussion on the "Turing test" provides another perspective of the limit of the intelligent machine. It provides an insight into judging the capabilities of technology, and how different humans can have very different perspectives and come to quite diverse conclusions over the same data set. In this paper, authors consider the capabilities of humans when it comes to judging conversational abilities, as to whether they are conversing with a human or a machine. In particular, the issue in question is the importance of human judges interrogating in practical Turing tests. The main point the authors consider here is the fallibility of humans in deciding whether they are conversing with a machine or a human, and hence they are concerned specifically with the decision-making process. A valuable feature of Turing's imitation game is not whether a machine gives a correct or incorrect response or a truthful or untruthful one, but rather whether it gives the sort of response that a human would give. Based on selected set of transcripts of an experiment on Turing test on conversation judgment, the discussion concludes that there is a long way to go before machines can be made to achieve human-like conversation for long periods of time and be able to enforce a considerable portion of a jury of human interrogators into believing that they were interacting with another human.

Aagaard's discussion on "absent presence" explores the influence of digital devices on everyday social interaction that can be visualised in terms of so-called absent presence, the state where a partner is physically present, yet absorbed by a technologically mediated world of elsewhere. The micro-social dynamics at stake in such impaired social interaction include delayed responses, mechanical intonation, a motionless body and a lack of eye contact. This type of impaired social interaction, the author says, leads to a mismatch between the vitality of a person and his or her absently present conversational partner. This amounts to a kind of unintentional mis-attunement which disrupts the smooth flow of ordinary interaction and signals indifference to what is being said. Taken together, these dynamics directly reveal the inattentiveness of absently present conversational partners. On this basis, absent presence is distinguished from related concepts of daydreaming and mind wandering. Although absently present persons may not actually be uninterested in what they are being told, it very much "seems like it". What we communicate through absent presence, in other words, is that we are uninterested in what our conversational partner is trying to tell us. Absent presence signals indifference to what is being said. This makes it all the more worrying when students report frequently experiencing a distinct lack of receptivity from friends and relatives who use mobile devices. Perhaps this is why so many people report having better conversations and higher levels of empathy when mobile devices are absent. In pace with the ubiquity of mobile devices and, by extension, absent presence, the experience may become so normalised that its upsetting element gradually wanes.

Tripathi's discussion on "Human-Technology Interaction" emphasises the embodied nature of communication, a need for the development of a phenomenology of technology, an inspiration for a new interface paradigm. It argues that in order for human users to share phenomenological experiences through multimodal systems, they need to deal with embedded computers, take into account of the embodied nature of our interactions with each other and the objects we manipulate, as ready-to-hand participants and tools. As technologies mediate more and more and transform human experiences with the world, they affect the embodiment of a user through human–computer interaction. This embodied transformation is now shifting boundaries between computers and everyday world. We see the world as we make the world, and we make the world into what we have seen and imagined through the tangible construction of technical possibilities. Any human act that makes a mark on any "outside" world also makes that world an extension of the human being who guides the change.

Contributing to the broader debate on the impact of technology on society, Ponse et al.'s discussion on "technology transfer motivation" in university learning environments gives a glimpse of how stimulus signals of positive motivating factors may contribute to getting the correct responses for improving the technology transfer process. Such positive factors include academic prestige, competition, generation of resources, the solution of complex problems, professional challenge, personal gains, personal gratification and the solution of society problems. The negative motivation factors include innovation environment, time required, lack of incentive and fear of contravening university policies. What is important in developing motivating environments is how to deal with the observer perception and the ways in which choices are made. It is proposed that inclusion of specific cultural aspects could provide more information about effective motivation factors that improve the technology transfer process. Naji and Ramdani propose a fuzzy logic procedure for learner's assessment. The procedure allows for the evaluation of learner environment by combining learner's responses and the degree of certainty of these responses. The authors suggest that this diagnostic procedure improves the process of content adaptation and self-adjustment on the one hand and makes the knowledge model clearly interpretable and more understandable to the learner and the teaching community.

Bhasin and Mehta's discussion on "The Diploid Genetic Algorithms" posits the need for the deployment of the premise of the dominance relation and that of the biological basis of evaluating dominance in developing robust and efficient machine learning techniques. Park's discussion on "Finnish Science Parks" illustrates how Technopolis Plc could grow rapidly particularly under the economic crises and explores growth strategies that could be implemented. Moreover, it also analyses the future prospects of growth and whether its growth strategies can be sustainable or not.

Ashrafi and Naghizadeh's discussion on "Architectural Works in the Achaemenid Era" makes us pay attention to the notion of the formation of culture and civilisation. The discussion explores the movement from idea to work, that makes the act of creation, and from work to idea that is concerned with the sphere of thinking, cognition and perception, thereby leading to the formation of culture and civilisation. It is held that the formation of any civilisation is a direct result of interaction between ideas and forms. If there is no idea, there will be no form, and if there is no form, promotion of idea would be meaningless. From the ancient time, the religious-mythical idea, as a whole idea, has been leading to other ideas including philosophical, mystical and political ideas. The authors assert that architecture, as a historical form, has been the outcome of interactions between these triple ideas. Knowing that historical form is established by interaction of ideas, the

political idea was supposed to be guided by knowledge splendour or philosophy rather than any other ideas, owing to the religious tolerance of the Achaemenids. Knowledge exchange between different nations was one of the main objectives of the Achaemenid, inspired by an ascending adaptation to the works of different nations. Cultural tolerance was widely expressed in this period as the zeitgeist and it brought up a kind of pluralism. Therefore, the architecture was expressed as an imperfect whole and perfect component, because the whole follows the plurality, while the component follows the unity (cross form), and, at this point, philosophical idea dominates mystical idea.

As the debate on artificial super-intelligence highlights the need for an ongoing conversation between technology and society, it also poses a challenge of the design of intelligent interactive and collaborative tools and systems to facilitate this conversation. AI&Society with roots in human–machine symbiosis provides a stimulating forum for debates on the AI machine. In its tradition of the widening and enriching the scope of AI and Society debates, the Journal has launched two new forums: a Student Forum and Curmudgeon Corner. The student forum provides a forum for young researchers to communicate their ongoing research to the wider academic community. Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. In response to the debate on AI and existential risk, Curmudgeon asks: Can we trust machine learning, when we do not understand what is going on in the black box? Can we and should we trust algorithms to cope with unanticipated errors and eventualities, even if we are aware that they are performing robustly at an anticipated minimum level of reliability? At a more fundamental level: What do we mean by an intelligent AI agent? This raises the question: What is intelligence and this in turn raises the question: What is human intelligence? Suppose we are aware of what intelligence is, a further question arises: How should an AI agent behave? And this raises the question: Do we know how humans should behave? When we make an ethical statement that AI should be developed for the benefit of society, a question arises as to what we mean by "benefit to society"? Ultimately the AI machine raises the question of values we hold and judgements we make about shaping the nature and path of technology. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? This question remains central to the on-going human-centred debate of AI&Society. AI&Society warmly welcomes contributions to the debate on the impact and implication of the super artificial intelligence on societies.

> It takes something more than intelligence to act intelligently.
> -Fyodor Dostoyevsky, Crime and Punishment

## References


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

Boyd R (2015) Man vs. machine: How humans are driving the next age of machine learning, CRUNCH NETWORK, Jun 11, 2015. http://techcrunch.com/2015/06/11/man-vs-machine-how-humans-are-driving-the-next-age-of-machine-learning/#.frrlobk:gSSP

CRASSH (2016) A symposium on technological displacement of white-collar employment: political and social implications. Wolfson Hall, Churchill College, Cambridge

Flint T, Turner P (2016) Enactive appropriation. AI Soc 31(1):41–49

Future of Humanity Institute (2013) Unprecedented technological risks, University of Oxford, Oxfrod, UK. https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf. Accessed 16 Jan 2016

Geist EM (2015) Is artificial intelligence really an existential threat to humanity? http://thebulletin.org/artificial-intelligence-really-existential-threat-humanity8577. Accessed 8 Jan 2016

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

Naughton J (2015) Should we be worried if our homes are soon smarter than we are? http://www.theguardian.com/commentisfree/2015/dec/06/smart-homes-security-risk-internet-of-things

The Responsible Data Forum (2015) Ways to practise responsible development data. https://responsibledata.io/ways-to-practise-responsible-development-data/. Accessed 5 Jan 2016