AI & SOCIETY (2020) 35:509–517
https://doi.org/10.1007/s00146-020-01035-6

EDITORIAL

Prediction paradigm: the human price of instrumentalism

Karamjit S. Gill[1]

Published online: 11 August 2020

Reflecting on the rise of instrumentalism, we learn how it has travelled across the academic boundary to the high-tech culture of Silicon Valley. At its core lies the prediction paradigm. Under the cloak of inevitability of technology, we are being offered the prediction paradigm as the technological dream of public safety, national security, fraud detection, and even disease control and diagnosis. For example, there are offers of facial recognition systems for predicting behaviour of citizens, offers of surveillance drones for 'biometric readings', **'**Predictive Policing' is offered as an effective tool to predict and reduce crime rates. A recent critical review of the prediction technology (Coalition for Critical Technology 2020), brings to our notice the discriminatory consequences of predicting "criminality" using biometric and/or criminal legal data. The review outlines the specific ways crime prediction technology reproduces, naturalizes and amplifies discriminatory outcomes, and why exclusively technical criteria are insufficient for evaluating their risks. We learn that neither predication architectures nor machine learning programs are neutral, they often uncritically inherit, accept and incorporate dominant cultural and belief systems, which are then normalised. For example, "Predictions" based on finding correlations between facial features and criminality are accepted as valid, interpreted as the product of intelligent and "objective" technical assessments. Furthermore, the data from predictive outcomes and recommendations are fed back into the system, thereby reproducing and confirming biased correlations. The consequence of this feedback loop, especially in facial recognition architectures, combined with a belief in "evidence based" diagnosis, is that it leads to 'widespread mischaracterizations of criminal justice data' that 'justifies the exclusion and repression of marginalized populations through the construction of "risky" or "deviant" profiles'. Prediction apps, such as the Australia's CovidSafe are now part of a wide variety of high-tech offerings to automate COVID-19 contact tracing. From Morrison (2020), we learn that prediction algorithms can be used to assess the outcome of patient X-rays and diagnose COVID-19 virus. For example, an Oxford-based data-visualisation company, Zegami, offers a machine learning model that quickly predicts the outcome of coronavirus patients by studying X-rays of their chests. However, AI algorithms need to be trained on a wider range of X-ray images from infected patients. Machine learning and data analytics are offered to 'accelerate solutions and minimize the impacts of the virus, and further machine learning tools are promoted to help expedite the drug development process, forecast infection rates, and help screen patients faster' in conjunction with the drug development process. However, these tools raise ethical issues of data protection, privacy, potential bias in the data or analysis, lack of transparency, explainability and accountability, and in the case of health care, it raises further questions of potential negative implications for the therapeutic alliance in patient–clinician relationships.

There is an argument that prediction tools can be used to bypass the messy biases and errors, for example in hiring managers by reviewing résumé data, ranking applicants and identifying top talent. However, Corinne Purtill (2020) notes that the machine learning hiring tool is only as smart as the input it gets. If sexism or other biases are present in the data, machines will learn and replicate them on a faster and bigger scale than humans could do alone. On the other hand, just as the tools can identify the subtle decisions that end up excluding people from employment, it can also spot those that lead to more diverse and inclusive workplaces. For example, Humu (Purtill 2020) uses artificial intelligence to analyze its clients' employee satisfaction, company culture, demographics, turnover and other factors, while its signature product, the "nudge engine," sends personalized emails to employees suggesting small behavioural changes (those are the nudges) that address identified problems. It is also the challenge of any organization attempting to nudge itself, bit by bit, toward something that looks like equity. Purtill quotes Iris Bohnet (a behavioral economist) that "The behavior is what matters, and the outcome is the same regardless of the

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

[1] University of Brighton, Brighton, UK

reason people give themselves for doing the behavior in the first place." We are thus back to the prediction paradigm. What matters is the purpose of behavioural prediction rather than just the tools of prediction. Recently it has dawned onto MIT that it makes no ethical sense to design prediction AI architectures that recognise 'people and objects in images' (Chadwick 2020) without also recognising the implication of 'assigning racist and misogynistic labels'. Although MIT has 'apologised for the 'racist and misogynistic' dataset', it is noted that "Despite this, apps and websites relying on neural networks that were trained using the database may spout out these shocking terms when analysing photos and camera footage." (Chadwick 2020). We learn from Prabhu and Birhane (2020) about how uncritical and ill-considered curation practices of large datasets pose threat to the identity and privacy of people and society. They cite 'ImageNet' as an example of the emergence of research culture that appropriates images of people as raw material without ethical scrutiny, and that of 'Clearview' as an example of 'secretive datasets' that 'currently exist hidden and guarded under the guise of proprietary assets?' They argue that although informed consent has been recognized as a critical component of big data including photographic data in domains such as medical and psychological sciences, there has been a wide spread erosion of 'the fundamentals of informed consent' and privacy. This wide spread erosion of ethics renders the claim of informed consent 'both ephemeral and vacuous'. The authors point out a deeply worrying and insidious threat of big data sets not only to vulnerable groups but also to the very meaning of privacy, as we know it. Although subtle forms of ethics such as those of '*ethics shopping*, *ethics bluewashing*, *ethics lobbying*, *ethics dumping*, and *ethics shirking*' are being promoted and propagated, they argue for a continued discussion of ethics and justice in machine learning to counter the danger of the appropriation of big data sets without ethical scrutiny. It is hoped that AI researchers engaged in facial recognition, machine learning models, and predication architectures would take serious note of the MIT example and cultivate an ethical and moral culture of designing AI tools and systems that are socially, culturally, ethnically sensitive, and strive to avoid harmful biases in surveillance architectures.

Harari (2020) argues that the epidemic of surveillance technologies that track, monitor and manipulate people, marks an important watershed in the history of surveillance. The danger lies in not just about the normalisation of the use/misuse of mass surveillance tools, it is the implication of a dramatic transition from "over the skin" to "under the skin" surveillance that has arrived with coronavirus. For example, what is now demanded of us is not just outside our skin but also inside—not just the blood pressure and temperature of our fingers, but also the blood pressure under the skin. He asks us to imagine a future scenario in which every citizen would be required to wear 'a biometric bracelet that monitors body temperature and heart rate 24 h a day'. The machine learning algorithms would know that we are sick even before we do, and 'they will also know where you have been, and who you have met.' It can be argued that the prediction architecture would not only shorten the chains of infection, but even cut the chain altogether, thereby could 'stop the epidemic in its tracks within days.' He cautions us about the implication of following this path of the prediction paradigm, when he says that cutting the chain of the epidemic may sound wonderful, but the danger is 'that this would give legitimacy to a terrifying new surveillance system.' In the euphoria of this surveillance-oriented digital future, it is crucial to remember that human behaviour, whether it appears in anger, joy, boredom and love, is a biological phenomena just like fever and a cough. Harari further says that the 'same technology that identifies coughs could also identify laugh, can also be used to harvest our biometric data en masse, not to 'just predict our feelings but also manipulate our feelings and sell us anything they want- be it a product or a politician'. This manipulation of biometric data, he says, 'would make Cambridge Analytica's data hacking tactics look like something from the Stone Age'. In response to this surveillance scenario, Harari proposes an alternative future in which instead of building a surveillance regime, we should aim to make use of new technologies to empower citizens, and use data to 'make more informed personal choices', and 'hold governments accountable for its decisions'.

To get an insight into the rise of instrumentalism we turn to Shoshana Zuboff (2019), who in her seminal book, *The age of surveillance capitalism,* provides a deep insight into the consequences of high-tech appropriation of the prediction paradigm. She warns that the goal of instrumentalism is to appropriate (or misappropriate) the prediction paradigm, not only to automate the transformation of our human experiences into behavioral surplus, a surplus resource for profit, but also to 'automate us'. We also learn how Silicon Valley misappropriated the affective computing architecture with the aim of automation of human emotion, the creation of an emotion chip, or the creation of emotion AI. In other words, not content with having automated the outer of the human self, that of human behavior, the high-tech aims to automate the inner being, emotion, and thus hollowing the body, treating it as any other object of profit calculation. The implications of the appropriation of these prediction architectures by high-tech are twofold: the first is the creation of a 'sense of inevitability of technology' and the second is the creation of the culture of technological dependency. The danger of this appropriation is that it adds a further layer to the culture of economic and market dependency and a sense of helplessness in the face of when the computer says "NO".

Zuboff (2019) notes that in 'ceding of our control of own bodies and behaviours', we were 'caught off guard' as we did not possess the lens of past experience to assess new threats and risks of the unprecedented onslaught of "instrumentarianism". Although it may be true about not having the lens to assess this onslaught, there was no dearth of the lens on past experiences of the danger of automation of the human dimension. A number of scientists, among them Joseph Weizenbaum, Hubert Dreyfus (USA), Mike Cooley (UK), and socially concerned movements such as Computers for Social Responsibility (USA), Human-centred movement (Europe), AI For Society conferences (UK) and AI & Society journal, were raising voices about the danger of instrumentalism as early as the 1970s and 1980s (Gill 1996; O'Neill et al. 2020). Joseph Weizenbaum (1976) alerted us to the limits of the universality of instrumental reason and the danger of its penetration into the culture of computation and machine learning. This, he argued, would lead to the reverence of the machine to the extent that human purpose is either ignored or misrepresented, as if every aspect of the real world can be formalised and represented in term of logical calculus. Mike Cooley warned us of the danger of automation of skill and knowledge, and argued for the development of socially useful technologies that cultivate and service the symbiotic relationship between the human and the machine. Howard Rosenbrock (Gill 1996), provided an inspiration for designing machines with purpose that enhance and facilitate the symbiotic expansion of both technology of purpose and societal benefits, in other words the enriching and expansion of both the tacit and objective knowledge. Inspired by Weiszenbaum's book, *Computer Power and Human Reason* (Weiszenbaum 1976), Mike Cooley's book, *Architect or Bee?* (Cooley 1987), and Hubert Dreyfus's book, *What Computers Can't Do* (Dreyfus 1978), AI For Society Conferences at University of Brighton provided a forum for socially useful artificial intelligence, as early as 1983. This laid the foundation for the AI & Society journal in 1986, which since its foundation has been a catalyst for a humanistic vision of art, science, technology and society. Given that these scientists and forums were raising voices about the danger of instrumentalism as early as the 1970s and 1980s, one wonders why these voices were not heard in countering the high-tech takeover of the prediction paradigm.

Could it be that the dominant computer science and AI research communities were very much absorbed and content with the narrow technical vision of progress. It may be that they were merely engaged in the design and promotion of machine automation, and unwittingly provided a camouflage for exploitation of their research, by high-tech giants Apple, Google, and Facebook, for profit. Could it be that high-tech companies were able to exploit the widening communication and service gap between on-time and online demands of the Society of Individuals and the capacity of the pre-digital era of public services to meet these demands. We learn from Zuboff that there emerged a gap between 'psychological yearning' of society for change, the yearning of individuals for just-in-time services and, in contrast, the indifference and incapacity of public institutions to meet these yearnings. The tech companies saw this gap and exploited this gap for profit. It is intriguing to note how high-tech companies highjacked the social concept of advocacy and appropriated it for profit—for example, Apple marketed its ipods and iphones as tools of emancipation, equality and inclusion. In other words, it was the technological advocacy and not social advocacy that offered emancipation of the individual—a new slogan of the digital society promoted by high-tech.

Although the AI community could not foresee the speed and motivation of high-tech to appropriate the predication paradigm, the community has recently been becoming more and more aware of the consequences of machine automation of the human, and the need to counter it. This we observe in an expanding interest in a humanistic vision of AI research among our authors. Not only does this vision counter the techno-centric paradigm, it also expands the symbiotic vision of technology and society. It builds upon Weizenbaum's ideas of the instrumental reason and 'judgment to calculation', to cultivate a broader vision of AI for society that facilitates engagement with a diversity of voices and over-the-horizon issues of arts, science, technology and society. Recent publications in AI & Society on themes such as social intelligence, robot ethics, philosophy of technological culture, streams of consciousness, cultural diversity and community technology design, bio-art, material hermeneutics and technoculture and technoscience, the dance of artificial alignment and ethics, and the trapping of AI agency, exemplify this vision.

In AI & Society circles, there has been growing concern that digital society is being promoted by not just high-tech but also by public policy makers without socially and culturally validated ethical, moral and legal constraints. The concern is not just the automation of behaviour and emotion but also the automation of behavioural interventions and modifications. The consequence of this automation is that it leads to the exclusion of human engagement to articulate, intervene and enforce ethical constraints on the misappropriation of predictive and affective computing architectures by the high-tech and its market forces for profit. Whilst in the 1980s, we faced the challenge of 'judgment to calculation', now in 2020s we face the challenge of the 'Human to Calculation'. It is no longer about the exclusion of the social but the exclusion of the human itself. It is crucial that we should not just be concerned with the ethical and alignment debates, we should also develop a strategic understanding into the process of the misappropriation of AI research by the high-tech for profit. Here the spread of the COVID-19 virus may provide us with a guide to this understanding. We

may pose the question, how did the spread of the prediction paradigm from academia to the high-tech market come about? To explore this question, two scholars come to mind: Shoshana Zuboff, whom we have met before in getting an insight into the appropriation culture of Silicon Valley, and Bruno Latour who in his recent interview on COVID-19, gives an insight into social networking of COVID-19 (Watts 2020). Reflecting on their insights, we can find that just like the COVID-19 virus, the prediction paradigm has not sprung from outside the human body but from within the academic body. It came from within the major AI research centres such as MIT and Stanford. Just like the virus, the prediction paradigm has spread from academia to Silicon Valley and then to the high-tech world like a tsunami. We now realise that the prediction paradigm could neither predict the COVID-19 tsumani, nor it could provide any relief or diagnosis to people who suffer from COVID-19. Just as the tsunami of virus cannot be controlled without human engagement and intervention (e.g. medical intervention, social distancing), the virus of the prediction paradigm cannot be controlled without social, ethical and moral constraints and interventions. It is worth repeating the argument that within the academic zones of MIT and Stanford, the prediction paradigm may have been constrained by ethical limits. But once it found its way to Silicon Valley, it was unconstrained by any ethical limits. Further, the COVID-19 pandemic has shown us that just as 'economy is a very narrow way of organising life and deciding who is important and who is not', so is making the digital future as our home a narrow technological way of thinking about what can be, what should be and what ought to be done for the benefit of society.

What we have also learnt from COVID-19 is that the spread of the virus crosses social, cultural, religious, ethnic and geographical boundaries, and thus can neither be controlled by these boundaries, nor can be abstracted away by quantification or wished or washed away through the technological narrative. So any attempt to externalise the spread of the virus to others or outside sources is not only shirking our social and ethical responsibility to mitigate its impact, but also harms others. Whilst we are in admiration of the humanistic spirit of medical, health and welfare professionals and carers in looking after COVID-19 patients and sufferers, the high-tech enterprises are offering machine learning systems for prediction and diagnosis of the virus. Whilst the medical profession is striving to make a case for our social and ethical responsibility to protect the spread of the virus not just to the self but also to the others, we are in danger of letting the high-tech shift this social focus to technical solutions, and in the process shifting ethical responsibility from the social domain to machine ethics in the technological domain. In the same vein, the spread of the virus and suffering of millions of people, families and communities cannot and should not be externalised to others, whether they are individuals, community groups or nations. We should all be inspired by the participatory spirit of medical and health care professionals, first respondents, and caring volunteers, in responding to the caring needs of those suffering the virus. This caring and collaborative spirit lies not in the abstract theoretical or methodological thesis of academia; it is deeply rooted in the professional, ethical and social responsibility ethos of the medical and care professions and welfare workers. We have learnt that when incorporating technological innovations such as those of data science and algorithmic tools in the diagnosis and treatment of patients, medical professionals and health care workers neither externalise their social responsibility to others, nor do they shirk in owning their social responsibility. This ethos of collaboration and responsibility are also being demonstrated by voluntary organisations, community groups and individuals in providing food banks, delivering food and medicine to the needy, and raising funds for health care, social welfare and survival needs of people and communities in many parts of the world. Again there is no thought of externalisation of social and ethical responsibility either to the other or to the outside. In the midst of the ethos of social and ethical responsibility and the spirit and practice of collaboration, we hope that this spirit would inspire the AI community to incorporate social and ethical responsibilities not just in the design and evaluation of AI systems, but also in the impact assessment of the algorithmic agency form a societal perspective. Such an impact assessment should include the prediction architecture of data science, machine learning, and deep learning, for example for ethical policy formulation, decision-making in societal domains, and the morality of embedding social robotics in health care environments.

Luis Moniz Pereira (2019), argues that the problem of the prediction paradigm is not that we have not lived with prediction, it is that we are in awe of the power of the machine, and are giving it too much power to automate human behaviour, without social, cultural, and legal, ethical and moral constraints. The idea that machine characteristics can be aligned with human values and morals seems to ignore the argument that the latter have evolved over centuries as a learnt behaviour, and cannot be just translated into logical rules. Pereira says that perhaps one day intelligent machines will live alongside humans, and through lived experience learn the norms of morality similar to ours. He further says that programming morality is a very complex problem, it has many dimensions. We are just starting to understand the challenges. It is like exploring a new continent. In addition to the articulation of the deep social changes and social instability triggered by the new robotic automation, we also need to cultivate a socially responsive practice to assess the implication of automating human behaviour and emotions. This should also focus on the enormous risks of social

instability and discontent inherent to the changes that are, and will be, caused by this automation. It is worth noting from Pereira (ibid.) that even the restrictive ethical and legislative proposal, ELLIS (European Lab for Learning and Intelligent Systems), risks delegating power to machine learning algorithms, void of moral standards. It ignores the 'notions of causality, of rule-based reasoning, of explanatory and justified support for decision-making choices, of arguing about ethical choices and exceptions'. This machine learning agency is based on the idea that systems can mine and learn from the huge volume of data, and thereby identify patterns of similarity to make decisions with minimal, if any, human intervention. If this bounded algorithmic agency lacks ethical constraints, then what makes us assured that the prediction paradigm can be tamed by ethical and moral constraints when it comes to the automation of human behaviour and emotion?

However, we also learn from the critical review (Coalition for Critical Technology op.cit.) that any move towards this taming of the prediction paradigm needs to 'embrace a historically grounded, process-driven approach to algorithmic justice, one that explicitly recognizes the active and crucial role that the data scientist (and the institution they're embedded in) plays in constructing meaning from data.' For this to happen, AI research community should cross-appropriate frameworks of situated practice and methodologies of grounded research from fields such as anthropology, sociology, media and communication studies, and science and technology studies. This also means that machine learning practitioners need to move beyond the dominant epistemology of abstraction and instrumental reason if they were incorporate societal concerns rather than excluding them from the design practice. By focusing on the technical vision of accuracy, precision and recall or sensitivity and specificity and performance metrics, the prediction paradigm perpetuates the narrow technical vision of progress.

AI & Society authors of this volume continue their own reflections on the ethical and alignment debates and making a contribution to the understanding of the prediction paradigm and its societal consequences. Coeckelbergh in 'Technoperformances' (this volume) reflects beyond the traditional tool use, medium and mediation relations of technology, and envisions technology, especially smart technologies as performing co-actors with humans. Drawing on performance metaphors from performance arts—dance, theatre, and music, he envisions human–technology interactive relations as techno-performances as if these were performed as co-choreography, co-direction, and co-conduct. It is this notion of co-action that transforms ethics and politics of technology into ethics and politics of performance. This view emerges from our engagement with the world, with technology, and with each other, giving us a more comprehensive view of what it means to live *with* technology and how our lives are increasingly organized *by* technology. The engagement and living with technology thus gives meanings to our experiences and actions. We wonder whether human relations with AI agency and its algorithmic systems and consequent debates on accountability, transparency and responsibility can be seen in terms of co-performances.

Tolga Yalur in 'Interperforming in AI' (this volume) offers a critical inquiry into the limits of the architecture of contemporary neural network models of machine learning that are applied in most commercial research such as Facebook AI Research. It provides an insight into the misemployment of 'natural' performance, and offers a 'context' as a variable of a performative approach, instead of a constant. It emphasises that the logic of performativity is not brought into account in all recurrent nets as an integral part of human performance and languaging. Moreover, recurrent network models also fail to grasp human performativity. The argument is that humans do not live in such a world; we live, perform, re-iterate and re-cite traces, and language is no exception. The ways we inhabit the world allows what Alan Turing calls imitation, and what Jacques Derrida calls repeating with differences. Each time we repeat a linguistic trace in the present, we do it differently from what we did in the past. The context, or space, revolves around the words and sentences in relation to the previous uses. NLP's language systems fail to grasp the contextual spirals, an integral part of human performance and languaging.

Curran et al. in 'Anthropomorphizing AlphaGo' (this volume) introduce us to cultural differences by both the Chinese and American press in their accounts of AlphaGo. The Chinese and American media's framing of Go thus offers a useful case study for consideration of these trends, as future advances in AI will are likely to bring humans into ever greater contact with "smart" machines. This and many other realms of human experiences will see both an influx of machine participation and digital interlocution. In their identification and explication of the framing of AlphaGo in the American and Chinese news media coverage of AlphaGo, the authors suggest that this framing is only a preliminary step in understanding the ways in which AI forces a dual re-evaluation of existing norms and attitudes about what constitutes the "human" and the "machine," and more importantly, what, if anything, fundamentally separates them.

Pedersen & Johansen in 'Behavioural Artificial Intelligence' (this volume) ask how do intelligent systems make inferences*?* And what insight do we have to ensure accountability the reliability, validity and accountability of judgments and decisions that artificial intelligent systems make? In their response to these questions, they argue that we should neither be content with developing artificial intelligent systems, focusing merely on the functionality of the systems, nor should the theoretical analysis be carried out detached from those who develop these systems.

Hayes et al. in 'Algorithms and Values in Justice and Security' (this volume) emphasise the notion of 'Value Sensitive Design' as another perspective of explicating accountability and transparency, accuracy, privacy, fairness and equality of algorithmic systems. They find that values are sensitive to disvalue if algorithms are designed, implemented or deployed inappropriately without sufficient consideration for their value impacts. This lack of value impact has the potential to result in problems of discrimination and constrained autonomy. The authors outline a framework of conceptual relations of values, and identify potential value tensions as a contribution to future research into value sensitive design of algorithms in justice and security. In raising concerns about the opacity and discriminatory aspects of algorithmic systems, they warn us of the danger of the uncritical acceptance of the process of data collection and its processing, when designing an accountability focused algorithmic agency. In recognising constraints of a value system design, they note that a failure to adequately incorporate values into the design and deployment process may not only be 'deleterious' to our values but also may actively inhibit their flourishing.

In their reflection on the changing nature of our social relations mediated by artificial intelligence (AI) assistants, Cunneen and Mullins in 'Artificial Intelligence Assistants and Risk' (this volume) argue that AI assistants present a significant societal risk amplified by their increasing use in healthcare, education, business, and service industry. To cultivate user risk awareness regarding AI assistants, the authors propose the creation of a risk narrative, which is focused on capturing, communicating and contextualising of risks of AI assistants, whilst supporting explainability as a risk mitigation mechanism.

In a response to gendered narratives of hopes and fears associated with intelligent machines and AI, Adams in 'Popularising Female Automata' (this volume), argues that a gendered reading of this narrative enables us to problematize the narratives associated with AI and expose the power asymmetries that lie within the association of technologies with traditional notions of femininity. In an effort to triangulate the schema of hopes and fears, Adams recognises that whilst such technologies represent the absolute dreams and human accomplishments, these also trigger a deep-seated fear that they will turn against us. The author argues that although the narratives of anthropomorphising of robots as the female are simultaneously a dehumanising of real women, these also provide an opportunity to reflect on the importance of narratives in shaping our current realities. For the author, the narrative realm itself too offers a space where the hopes and fears of AI can be fulfilled, and moreover, the ethical consequences of the fulfilment of these hopes and ideas can be played out and experienced. However, these narratives and their reflections should counter the robotic view that real women can be considered 'mere machines', whose affectations and even 'intelligence is merely simulation.'

In their review of research into algorithmic appreciation and algorithmic perceptions, Araujo et al. in 'AI we trust?' (this volume) explore the perception of risks and opinions of Dutch people about fairness and usefulness of automated decision-making at a societal level. The authors find that whilst privacy is considered a pivotal aspect, their rather optimistic findings somewhat contrast the rather critical and pessimistic tone that is often prevalent in media reporting as well as in the academic literature. This pessimistic tone highlights fears over bias, loss in human dignity and autonomy, and more generally concerns about 'AI taking over' and replacing human decision-makers. In their conclusions, the authors suggest that the Dutch population is at least 'not blind' to the potential benefits of automated decision-making (ADM), in terms of its usefulness and fairness, even though they do see risks. They further observe that humans as decision-makers are not per se perceived as being irreplaceable in comparative scenarios of decision-making in specific scenarios. The authors recognize that public perceptions about the potential usefulness and fairness of ADM are not the same as those of the individual and societal acceptance of actual automated decisions. They thus suggest a caution in interpreting these findings as a basis for government initiatives in the Netherlands, as well as, elsewhere in the EU to explore the potential of ADM in various sectors of society.

Goagoses et al. (this volume) provide an insight into challenges of research collaborations with indigenous communities, thereby making a contribution to an ongoing debate around appropriate ethical conduct of researchers in situ. They point out that the notion of indigenous communities still being in control of their information, even after they have shared it with the researcher, can be difficult to understand. A major concern is that researcher–participant interactions have a tendency to fall back on prior practices of "good" behavior, without much reference to contextual and cultural practices. It is noted that this tendency may be with the mismatch between level of familiarity with concepts and actions and the assurance for interpretation accuracy, specifically when aligned with known ethical guidelines and academic processes. Further, they suggest that mismatch between the expected behavior of participating communities and the novice researchers may lie in mis-understanding of the interaction guidelines and their translation into the intended behavior. The challenge also lies in dealing with contradiction between the theoretical understanding of interaction and the intent of being respectful towards the other culture, as well as, the ability of novice researchers to adopt a modified practice of respectful behavior. Another challenge is how to update and overwrite existing research practices that are rooted in deeply ingrained understanding

and familiarity of research practices. This requires a deeper and more guided engagement with the research guidelines.

In putting forward an argument for a future scenario of an artisanal economy, Ron Eglash et al. in 'Automation for the Artisanal Economy' (this volume) wonder whether AI, robotics and related automation technologies can enhance the economic viability and environmental sustainability of the beloved crafting professions, perhaps even expanding their niche to replace some job loss in other sectors. And further whether artisanal labor, combined with technology, could potentially help to democratize the economy, allowing independent, small scale businesses to flourish. In this exploration, the authors examine the possibilities of utilizing AI to support hybrid forms of human–machine production towards a future vision for more "generative" economic forms in which labor value, ecological value and social value can circulate without extraction or alienation.

Spicheva and Polyanskaya in 'Culture codes of scientific concepts' (this volume), introduce the reader to Rapaille's concept of culture codes, and Hall's encoding and decoding model of communication, to identify the culture codes of scientific concepts in global scientific online discourse. As an example, the authors attempt to identify the culture codes of the concept of "image", because this concept can be interpreted in different ways in Russian and international scientific discourse. It is noted that these interpretations are quite different, despite the fact that the scientific concept of the image is interpreted in both the technical and natural sciences, as well as, in the social and human sciences in the Russian national and international online discourse. They argue that the concept of image investigated in their research has quite different culture codes in the Russian national and international scientific online discourse. It is further argued that knowing such differences and similarities in interpretations of scientific concepts is important for the integration of national scientific research into the international scientific discourse. The researchers thus must keep in mind that quite different interpretations of scientific concepts exist, and take them into account, when they decide to engage in the international discourse. They suggest that their method may be used for revealing the culture codes of any scientific concept (using any citation database), which can contribute to revealing and understanding the interpretations of these concepts by researchers from different countries.

Polak et al. in 'Intelligent finance and treasury management' (this volume) argue that it is the virtual nature of functions and processes of finance/treasury that lend themselves to increasing automation. For example, at present, the AI network neural system is widely used in many fields of treasury management, such as early warning of potential financial crisis, diagnosis of financial risk, control of financial information data quality, and mining of hidden financial data, information. Artificial intelligence in finance and treasury, they argue, is thus most analogous to the complexity of a human nervous system as it encompasses far more than the automation of tasks. Similar to the human nervous system, AI systems in finance/treasury must manage data quickly and accurately, including the capture and classification of data and its integration into larger datasets.

Cohen and Regazzoni (this volume) present a machine learning virtual platform, a 'Leap Motion controller, for hand rehabilitation for post stroke patients.' The authors note that physicians and physiotherapists monitor and assess the improvement of patients' rehabilitation through a web application. They suggest that the proposed low-cost technology-assisted rehabilitation processes can be easily exploited at home. Sreelekha in 'Indowordnet' (this volume) draws our attention to the challenge of translating various lexical phenomena that Indian–Indian and English–Indian language machine translation (MT) system development faces, especially in handling the ambiguity during translation.

Neri and Cozman, in 'The Role of Experts….' (this volume) give an insight into the role of experts in creating public perception of the risk of AI that is mostly associated with existential risks. They suggest that the source of this perception resides in the public positioning of experts, who happen to be the real movers of the risk perception of AI, instead of actual disasters. The authors suggest that this perception was framed and communicated by a number of public intellectuals who promoted the recent idea that AI could be a real threat and endanger all humans. They further comment that message of risk was based on counterfactual scenarios instead of actual events, such as the crash of a particular self-driving car. They argue that this framing of risk ignores the possibility of trigger events that arise from sheer human conjectures. It doing so, this framing ignores the active role of authoritative individuals, such as the experts, in the social interplay of the public position, and what such positioning can bring to the expert. It is, however, recognized that for any technological development, experts may not have a consensual public position about the risk of such technology; but may display three different positions—they can be antagonists, pragmatists or neutrals, and enthusiastic experts. In the case of AI, the argument is that antagonists believe there are insurmountable barriers to achieve full-fledged, human level AI; so any risk scenario related to that is 'nonsensical'. Pragmatists or neutrals believe that it is hard to even depict what are the real challenges to develop a full-fledged human level AI; even though we may achieve it at some point. For this group, the real dangers are in the short-term and are related to the portion of technology that already works in the world, such as the effect of biased datasets for machine learning algorithms. Finally, the enthusiastic experts believe that the full development is just a matter of time, and such a development

will bring a profound change. However, changes can be positive or negative. Because of that, enthusiastic experts can be grouped into pessimists and optimists. The authors argue that existential risk scenarios are framed by the pessimists. Further, they argue that those pessimist experts who happen to be risk communicators, and are capable of amplifying their messages based purely on the conception of such counterfactual scenarios, they can trigger many indirect effects within society.

On the regulation of autonomous mechanisms, Cristianini and Scantamburlo in 'On Social Machines for Algorithmic Regulation' (this volume), discuss the possible implication for human autonomy and social order, of building blocks of algorithmic regulation that are already well in place. Building on the notion of the 'social machine', the authors identify convergent social and technical trends that are leading towards social regulation by algorithms, and reflect on their possible social, political, and ethical consequences. They observe that although the algorithmic regulation of society may be of little more than a tempting idea in academic, policy and entrepreneurial circles, many of its components already exist. The authors alert us to a political interest in deploying some version of this regulation in the form of surveillance. Taking ORCID numbers as a scoring criterion of journal articles, the authors illustrate how this academic scoring system may lead to an unintended drift of academics towards conforming to a behaviour dictated by the scoring system as if it were another social platform of interaction. Although there may be a scope of 'opting in' or opting out' of such social platforms, it is asserted that this form of "*gravitational pull*" (e.g. algorithmic regulation of society) exerts a force that brings ever larger portions of people's lives into them. As the system scales up, the authors argue, the cost of opting-out increases with the size (or coverage) of such a system. Not only is this used in viral marketing strategies, but also this creates a 'Nash equilibrium' where everyone is part of the system: at that point there is a significant cost for each individual to leave. They wonder whether a business today can afford not being on the internet. And is it still meaningful to claim that people have freely opted into such a system? In posing these questions, the authors lay out the essence of the dilemmas of persuasive technologies and algorithmic regulation. The authors make a plea that scholars in Ethics, Sociology and Engineering may find a way to jointly address these questions.

Hoffmann and Hahn in 'Decentered ethics' (this volume) propose a policy framework for ethics that is informed by a sound philosophical underpinning. In the pursuit of an ethical framework, they reflect on the nature of ethical AI systems. This reflection focuses on the role of moral agency and 'patiency', the implication of vague and ambiguous concepts or the problem of other minds, and the impact of our philosophical and conceptual analysis on the regulatory landscape.

AI & Society, rooted in the humanistic tradition of art, science, technology and society, continues its own evolution responding to the ongoing debates on technology and society. During 1980s, its focus was on machine intelligence, followed by its focus on human-centred systems during 1990s, and at the beginning of the twenty-first century its focus shifted to knowledge, culture and communication. Now the challenge for AI & Society is to respond to the prediction paradigm and its consequential misappropriation by the high-tech market for profit. This response should include ethical and moral challenges of surveillance architectures that are being promoted as a technological panacea of societal problems, including those made visible by the COVID-19. In reflecting on these challenges we quote Mike Cooley (the author of *Architect and Bee*?) who crystallises the essence of AI & Society when he says that: 'the future is not "out there" in the sense that a coastline is out there before somebody goes to discover it. It has yet to be built by humans'. So, it is up to AI scientists, engineers and practitioners to be architects of the technological future and NOT let the Silicon Valleys discover this future for us.

References

Chadwick J (2020) MIT apologises after a giant dataset it was using to teach AI how to recognise people and objects in images was found to be assigning racist and misogynistic labels. https://www.dailymail.co.uk/sciencetech/article-8483929/MIT-pulls-racist-misogynistic-dataset-offline.html. Accessed 4 July 2020

Coalition for Critical Technology (2020) Abolish the #TechToPrison-Pipeline Crime prediction technology reproduces injustices and causes real harm https://mail.yahoo.com/d/folders/1/messages/AHGH7L96Q_7wXvJV0Qu_0LlmGls?reason=invalid_cred. Accessed 24 June 2020

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

Dreyfus HL (1978) What computers can't do: the limits of artificial intelligence. HarperCollins, New York **(Revised, Subsequent edition)**

Gill KS (ed) (1996) Human machine symbiosis: the foundation of human-centred design. Springer, London

Harari YN (2020) The world after coronavirus, Financial Times, 20 March 2020. https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75?fbclid=IwAR1dkuffhX1iglJFL9ERKvyXRaOSkK7qE_9NOlXZJWvQQw3aHyrXIRRn1vU. Accessed 20 June 2020

Morrison R (2020) Artificial intelligence that can diagnose COVID-19 using X-RAYS could help identify cases of coronavirus more quickly and predict outcomes for patients, computer programmers claim. Mailonline 6 April 2020 https://www.dailymail.co.uk/sciencetech/article-8191549/Artificial-Intelligence-help-diagnose-COVID-19-using-X-RAYS.html. Accessed 20 June 2020

O'Neill B, Stapleton L, Gill KS, Brandt D (2020) A Discourse on AI and society: your calculus may be greater than his calculus. But will it pass the Sullenberger Hudson River test? https://vimeo.com/433640813/4c7f85f902. Accessed 28 June 2020

Pereira LM (2019) Should I kill or rather not? AI & Soc 34:939–943. https://doi.org/10.1007/s00146-018-0850-8

Prabhu VU, Birhane A (2020) https://arxiv.org/pdf/2006.16923.pdf. Accessed 4 July 2020

Purtill C (2020) Algorithms learn our workplace biases. Can they help us unlearn them? New York Times, March 10, 2020. https://www.nytimes.com/2020/03/10/us/algorithms-learn-our-workplace-biases-can-they-help-us-unlearn-them.html. Accessed 20 June 2020

Watts J (2020) Interview with Bruno Latour: 'This is a global catastrophe that has come from within'. The Observer. Sat 6 Jun 2020. https://www.theguardian.com/world/2020/jun/06/bruno-latour-coronavirus-gaia-hypothesis-climate-crisis. Accessed 15 June 2020

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

Zuboff S (2019) The age of surveillance capitalism. Profile Books, London