Journal of Risk Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjrr20

Ignorance and the regulation of artificial intelligence

James M. White & Rolf Lidskog

To cite this article: James M. White & Rolf Lidskog (2022) Ignorance and the regulation of artificial intelligence, Journal of Risk Research, 25:4, 488-500, DOI: 10.1080/13669877.2021.1957985

To link to this article: https://doi.org/10.1080/13669877.2021.1957985

JOURNAL OF RISK RESEARCH
2022, VOL. 25, NO. 4, 488–500
https://doi.org/10.1080/13669877.2021.1957985

Ignorance and the regulation of artificial intelligence

James M. White and Rolf Lidskog

Environmental Sociology Section, School of Humanities, Education and Social Sciences, Örebro University, Örebro, Sweden

ABSTRACT

Much has been written about the risks posed by artificial intelligence (AI). This article is interested not only in what is known about these risks, but what remains unknown and how that unknowing is and should be approached. By reviewing and expanding on the scientific literature, it explores how social knowledge contributes to the understanding of AI and its regulatory challenges. The analysis is conducted in three steps. First, the article investigates risks associated with AI and shows how social scientists have challenged technically-oriented approaches that treat the social instrumentally. It then identifies the invisible and visible characteristics of AI, and argues that not only is it hard for outsiders to comprehend risks attached to the technology, but also for developers and researchers. Finally, it asserts the need to better recognise ignorance of AI, and explores what this means for how their risks are handled. The article concludes by stressing that proper regulation demands not only independent social knowledge about the pervasiveness, economic embeddedness and fragmented regulation of AI, but a social non-knowledge that is attuned to its complexity, and inhuman and incomprehensible behaviour. In properly allowing for ignorance of its social implications, the regulation of AI can proceed in a more modest, situated, plural and ultimately robust manner.

ARTICLE HISTORY
Received 13 October 2020
Accepted 7 July 2021

KEYWORDS
Artificial intelligence; risk regulation; ignorance; non-knowledge

Introduction

Attention to artificial intelligence (AI) has increased dramatically over the past decade. There are great hopes and expectations for what these technologies can achieve, and both the public and private sector have made substantial investments in their research, development and application. But along with their well-publicised benefits come new worries and regulatory challenges: that these technologies will be intentionally put to destructive ends, or that they will have unintended and severe consequences. Thus, entwined with the development and roll-out of AI, an ecosystem of ideas, practices, experts and organisations is emerging to respond to the technical assessment and public perception of the risks that they pose.

While the direction and distribution of AI will be affected by these hopes and worries, it is important to remember that AI is an existing technology that has already thoroughly permeated and changed society and social life. Not only does AI make certain activities faster and more efficient, it also affects them qualitatively, thereby gradually and often invisibly reshaping social

CONTACT James M. White james.white@oru.se Environmental Sociology Section, School of Humanities, Education and Social Sciences, Örebro University, Örebro, Sweden.

relations, practices and institutions. Society is not only using these technologies but becoming dependent on and even partly constituted by them (Kröger 2021).

Social scientists have played no small role in producing and disseminating knowledge that grapples with the social impact of AI. Considerable research has been published on the social consequences of AI and related technologies (e.g., Eubanks 2018; Stilgoe 2018), as well as efforts to minimise the harms that they can cause (e.g., Scherer 2016; Calo 2017).

In this paper, we foreground the other side of social knowledge, that is, social ignorance, and its importance to the regulation of AI. By this we do not mean the ignorance of AI to the linguistic and social worlds inhabited by humans, although that is certainly important (Collins 2021). What we are instead interested in is the ignorance of scientists and policy-makers to AI systems, their behaviour and what they mean for society. This ignorance is not something that can be avoided or denied. It is not simply a deficit that needs to be overcome. Rather, ignorance is a necessary consequence of the situatedness of knowledge holders (Alcoff 2007). Following Gross and McGoey (2015), we stress that ignorance is regular to all decisions about how to regulate and investigate AI. Without recognition of what is unknown about these technologies both acknowledgement of the risks that they pose and the ability to steer their development are hindered. New insights in turn raise new questions and a greater awareness of what remains uncertain and unknown. We use 'ignorance' in a general way to name what lies beyond the limits of knowledge. The concept of 'non-knowledge' we use more specifically to refer to efforts to understand ignorance to guide future action (see Gross 2007). The contribution of the paper is to discuss AI risks and their regulation from this perspective to inform the way that social knowledge about them is produced.

Within the broader public discourse, AI acts as an umbrella concept that refers to everything from rather weak AI (e.g., a computer game) to strong AI (e.g., the belief in a future 'superintelligence'). In computer science, a distinction is often made between symbolic AI (such as expert systems), in which the developer fully specifies the objects and relations known to a system, and subsymbolic AI (i.e., self-learning algorithms, such as artificial neural networks), in which computer models are trained on large, labelled datasets (Russell and Norvig 2016). As it has been the driver of recent interest and investment, our use of the term refers, in the first instance, to the latter of these.

Having said this, we also assert that AI cannot be thought of as simple, closed technological systems. Even at a function level, they must be regarded as complex, open, sociotechnical systems that rely on and interact with broader material infrastructures, and social, political and economic institutions and organisations (Lindgren and Holmström 2020). This conceptualisation challenges many of the implicit assumptions of what AI is and how it will affect society, and broadens the remit of what AI research can and should involve. In much of the popular, technical and industry literature on the effects of AI, it is assumed that social change will be driven by technological innovation. Our approach challenges this narrative by situating technologies, technologists and firms within longer discursive, cultural and political economic trends, in which they no longer play an all-encompassing role.

Accordingly, the risks presented by AI are understood to be social problems in need of social as much as technical solutions. This means moving away from narrowly-defined mathematical or technical risk mitigation, suspending identification of the public's apparent misperception of risk, and being wary of too strong an emphasis on risk communication (Lidskog and Sundqvist 2012). It means recognising that the risks associated with AI have complex and often indeterminate causes, and that there are limits to the ways that these have been framed within the literature. It means understanding risk as a relational process that is affected by both social practices and institutions, and material and discursive forces. It means stressing the systemic character of risk and acknowledging that risks are impossible to discover, understand and regulate without considering their wider context (Renn 2021; Schweizer 2021). Rather than conceive of risk in terms of the probabilities of an undesirable

event, we approach them through the frame of what is known and unknown about their character.

By focusing on ignorance, this paper explores how social knowledge and non-knowledge can contribute to a better understanding of AI and the regulatory challenge that it entails. This is done by reviewing literature on AI and associated risks, including contributions both from within the community (i.e., researchers related to AI research) and external evaluations (i.e., mainly social scientists). Our argument is made in three steps. First, we investigate risks associated with AI and the attendant quest for social knowledge on how to handle them. These risks are often technically framed, leading to a narrow and instrumental role for social knowledge. The problems of transparency, accountability and bias illustrate how social scientists have sought to reposition the political and social condition of AI risks. Second, we turn to the material characteristics of AI and the broader panorama of risks that they pose to society. Central here is the epistemic claim that AI cannot be fully known; their complexity, and inhuman and incomprehensible behaviour inhibit understanding not only for outsiders (e.g., citizens and regulators) but also for insiders (i.e., researchers and developers). Enacted by a fragmented regulation, AI is deployed in complex and pervasive ways, resulting in impacts and risks at various scales that are almost impossible to overview. In the final step, we return to the issue of social knowledge in AI and its regulation. We argue that better taking non-knowledge into consideration will lead to improved regulation with fewer unintended consequences and failures. Thus, by better recognising the limits of AI, and the limits of what we know and can hope to know about their risks, social science can make more substantial contribution to the regulation of AI.

Social knowledge of AI risks

Alongside the recent resurgence of AI, there has been a rise in public perception of AI risks (Neri and Cozman 2020), and academic literature promoting safer and fairer AI (e.g., Scherer 2016; Boddington 2017; Corbett-Davies and Goel 2018; Benjamin 2019). This interest in risk and risk management varies from the mundane to the speculative. Philosopher Nick Bostrom (2014), for example, has achieved widespread acclaim for his exploration of a hypothetical 'superintelligence'. The AI that we are interested in is more familiar than this, and it would be a mistake not to recognise how it overlaps with big data, self-learning algorithms and autonomous robots—each of which is an established topic of concern for social scientific research. In this section, we explore risks associated with these technologies and highlight the important contributions that social knowledge has made, both in industry and in the literature.

Many computer and data scientists have come to realise that data is often biased and that algorithms can sometimes cause harm. In 2015, Google image tagging software labelled an African American couple as gorillas, which is not only offensive but an example of the poor performance of the model on images of non-white faces (Benjamin 2019). In 2016, the newsroom *Propublica* exposed the racial inequalities of predictive policing systems (Angwin et al. 2016), which lead to a flurry of articles on the challenges of correcting data biases and emphasising fair outcomes (Broussard 2018). In 2017, a paper in *Science* showed that self-learning word association algorithms can produce race and gender stereotypes (Caliskan, Bryson, and Narayanan 2017). Bolukbasi et al. (2016) describe a Google Natural-Language Programming model that adds the final word to the sentence: "Man is to computer programmer as woman is to homemaker". And in 2018, an Uber autonomous vehicle struck and killed a pedestrian in Tempe, Arizona, raising concerns of cause, agency and accountability in mixed human–machine systems (Elish 2019).

Industry and academia have responded to these high-profile events by founding new areas of study, such as Fair Machine Learning, and annual meetings, such as the well-attended Conference on Fairness, Accountability, and Transparency (Corbett-Davies and Goel 2018). While important and commendable, these efforts are geared towards formal, technical solutions to what are often complex social problems with long and deeply entrenched histories (see Eubanks 2018; Noble 2018). To pick one example, a paper by Rahwan (2018) argues for programming the social contract into an AI system. This re-imagines the 'human-in-the-loop' control concept as a 'society-in-the-loop', which the author summarises with the equation: 'SITL = HITL + Social Contract'. How the social contract would be translated into code is a problem left to others.

There is now a recognition that the social sciences have much to offer to improve the regulation and roll-out of AI. A leading AI textbook (Russell and Norvig 2016) recognises the contributions of philosophy, economics, psychology and linguistics to the development of the field. Beyond this, industry has shown its support for research with a multidisciplinary agenda. For example, Intel employ in-house anthropologists, such as Genevieve Bell, and Microsoft and Google are supporters of The AI Now Institute, founded by Kate Crawford and Meredith Whittaker. There are several leading social science research institutes focusing on AI and related topics in the UK (such as the Oxford Internet Institute and the Leverhulme Centre for the Future of Intelligence), and in Sweden, The Wallenberg AI, Autonomous Systems and Software Program have established a strand for the social sciences and humanities, the director of which has argued for AI to be a rich, interdisciplinary field (Dignum 2019).

While social science can pluralise and legitimise AI development and regulation, and help ease public worries, the contributions made by social scientists have far exceeded these instrumental ends. Online and in the academic literature, they have emphasised the social nature of AI risks and have pushed back against the drive for technical solutions. This is evident in debates on transparency, accountability and bias.

The transparency problem in AI is often regarded as one of the major obstacles to effective risk governance. However, discussions of the concept tend to refer to a number of different things (Burrell 2016), such as: an intentional feature of the system (e.g., for security purposes); the secrecy inherent to competitive markets (Pasquale 2015); industry coding practices, including the use of metaphors (Larsson 2017); the level of technical expertise required to understand software languages and computer programs; the complexity of deployed software systems (Kitchin 2017); and the incomprehensibility of trained neural networks (Mittelstadt et al. 2016). While efforts have been made to address some of these issues through interpretability, explainability and justification in algorithmic design, improving the reporting of systems alone does not address the social issues or pressures that prefigure them. Ananny and Crawford (2018) argue that narrow definitions of transparency ignore the broader assemblage through which algorithms perform work. Transparency, they suggest, is a distraction from the social, political and economic context of algorithmic decision-making. Reddy, Cakici, and Ballestero (2019) go further, arguing that even solving the transparency problem in machine learning would not fully account for issues of accountability.

Knowing who to blame and who to hold responsible in the event of AI failure is a thorny issue. While moral philosophy's notorious shopping trolley problem is sometimes applied to autonomous vehicles as a paradigmatic example of AI ethics, this abstraction neither fully anticipates the complex issues that AI are likely to meet outside the laboratory, nor absolves programmers of responsibility for the choices that they make during a system's development (JafariNaimi 2018). The main technical problem is that the agency of AI make them wholly unlike other advanced technologies (Boddington 2017). Indeed, empirical research shows that AI professionals are unwilling to accept full responsibility for the systems that they work on, and instead present themselves as mediators between the agencies of users, the technology and regulators (Orr and Davis 2020). While it may be desirable to curtail machine agency by placing humans in- or on-the-loop, it is unrealistic to expect that a human controller will be

fast enough to take over from misbehaving machines in all cases (Morgan, Alford, and Parkhurst 2016). The role of humans within such a system may be, as Elish (2019) has suggested, to act as 'moral crumple zones', that is, to bear the brunt of social and legal blame.

In a similar vein, Crawford (2017) has argued that research on machine learning bias focuses too much on what she calls 'harms of allocation' and not enough on 'harms of representation'. The first of these refers to the distributive outcome of an unequal AI system. These are immediate, easily quantifiable, discrete and transactional, and therefore possible (if not always easy) to address at a technical level. In the case of the predictive policing algorithm, fairer outcomes might be achieved by pre-processing data to account for race (although race is often excluded from databases for privacy reasons, see Whittaker et al. 2018) or post-processing results to ensure fairer outcomes. More difficult to address, however, are the second type of harms. Harms of representation refers to the ways that social identities can be portrayed by an AI system. They are long-term, difficult to formalise, diffuse and cultural. In terms of the Google image tagging failure discussed above, while it may be possible to 'fix' the model by removing the gorilla tag or by properly sensitising the costs of misclassification (Russell 2019), this does not contend with what it means to call a black person a gorilla, or how such language is interwoven with histories of race science and the dehumanisation of ethnic groups (Benjamin 2019). Framing bias in terms of potential harms highlights the importance of social (and not only technical) approaches to the anticipation and management of AI risk.

In debates on transparency, accountability and bias, social scientists have shown the limits of technically-oriented, closed systems theorising. Too much attention to the transparency problem downplays the political economic context in which AI are developed and are able to do work. On the issue of accountability, several social scientists have asserted that developers should not be allowed escape responsibility when something goes wrong; that blame cannot only be attributed to the agency of the AI or the inattention of its human operator. Finally, while research on machine bias serves an important function, harms related to the representation of social groups have a history to which there can be no technical fix. These contributions have helped bring nuance and legitimacy to AI, and have improved knowledge of the risks that they pose. But they have also been met with push-back from industry. Before exploring what this means for social knowledge in AI, we re-classify the characteristics of AI according to their visibilities as a topic of study.

The (in)visibilities of AI

In this section, we are less interested in the contribution that social science has made to AI discourse than the materiality of AI systems themselves. In a break with terms usually used to describe them—the limits of which were discussed above—we characterise AI in terms of their invisibilities and visibilities.

Visibility is used to refer to the ways that AI reveal themselves, especially to scientific enquiry (following Larkin 2013). This is not as straight-forward as it might seem. AI can be tough to grasp as an object of theoretical and empirical analysis. The term is used in various and sometimes broad ways, and its essence or limits are difficult to specify. This is a familiar problem to AI researchers that is often connected to the conceptual ambiguity of 'intelligence' and the tendency for the goal posts to shift whenever AI researchers manage to line them up. The broad definition of AI that we have adopted, which invites reflection on distributed agency and responsibility, further complicates the term. Despite these challenges, it is possible to recognise some general characteristics of AI systems. We begin with the ways that AI escapes notice.

As an area of expertise, AI is fast-moving and opaque to outsiders. Most people do not have the technical knowledge required to train neural networks. While simple models can be built with an elementary knowledge of Python (see Broussard 2018), more complex models require

advanced skill and expertise. There is, for example, considerable variation in the salaries of newly graduated developers. An employee of Microsoft said during a 2017 lecture: "If a kid knows how to train five layers of neural networks, the kid can demand five figures. If the kid knows how to train fifty layers, the kid can demand seven figures" (quoted in Mitchell 2019, 112).

But even well-paid and highly-experienced engineers and software teams cannot know everything about the system in which their code is deployed. Large software projects are composed of many small functions and algorithms, which originate from different libraries or frameworks, and are structured in complicated and often recursively nested decision trees (Kitchin 2017). Beyond the project code itself there is a broader infrastructure of hardware, software, data and communication protocols that must be trusted and relied upon. Beyond this: the complex human world of social norms and behaviours.

Not only are AI software and systems complex in design, their ability to learn also means that their behaviour is unknowable and at times unexpected. Neural networks consist of many layers of artificial neurons that become sensitised to specific data patterns during training. As they are unable to report back to a human operator about these patterns, they can be known only by observing their response to new input (i.e., they are a black box). Put differently, not only do AI apprehend the world in inhuman ways, but there is no easy translation between their ways of knowing and ours. For example, while they are capable of powerful image recognition and classification, AI may not 'see' what their developers want them to see. Mitchell (2019) describes an attempt by one of her students to create an AI that is able to distinguish photos of animals from photos without animals. Rather than a generic animal outline, the model fixated on the blurry backgrounds that photographers often use to bring their subject into focus. Similarly, while the best image recognition models can distinguish between many different species of birds, animals and plants, when they fail they often do so (from a human perspective) in a spectacular manner. Researchers have shown that image recognition models can be fooled by making minor changes to an image's pixels that a human would fail to notice (Szegedy et al. 2014), or by making images of near-random pixels that to a human look like static, but which the model classifies with a high degree of confidence (Nguyen, Yosinski, and Clune 2015).

AI models deployed outside the laboratory will encounter situations that they have not been specifically trained to handle. Even if they act appropriately in most cases, the ways in which they fail will be unpredictable and potentially catastrophic. Evoking the concept of the long tail used in business and economics, Sendhil Mullainathan refers to the inevitability of AI failure in unanticipated real-world encounters as 'tail risks' (see Lanier 2020). Mitchell (2019) develops this idea, arguing that the brittleness of AI models and the lack of research on how they fail are areas of potential harm that have received insufficient attention from regulators. Given the varied and complex ways that AI are being deployed, such unpredictability could amplify other risks or produce devastating cascade failures. And once something does go wrong it is not certain that the problem could be addressed. Because the behaviour of AI models is incomprehensible to humans, it may prove impossible to fully know and correct the cause of the initial failure.

Not all of the characteristics of AI are so difficult to perceive. It is uncontroversial to observe, for example, that the technologies are already in widespread use. Anybody with an internet-enabled smart phone has access to powerful AI models, which can be used, for example, to recognise and interpret a spoken statement, identify an appropriate nearby restaurant, and translate a menu into a different language. But even without directly calling upon these systems, a phone user is still enrolled by them. Smart phone apps, operating systems and network operators continually generate data through a person's phone (passively, actively and through voluntary submission), including information about its use, location and surroundings (Kitchin 2014). Some of this is data is uploaded to servers, before being cleaned, processed (in some instances anonymised) and stored, possibly with the intention of selling it to a third-party broker. Not having a phone is no guard against this enrolment. Increasingly, the cards in our wallet, the devices

and appliances in our home, the vehicle we drive, the public transport we take, the way that we move through public and private spaces, and our interactions with companies and the state are all part of a system to track, sense, measure and otherwise turn our everyday lives into data (Zuboff 2019). This data, in turn, is used to train new AI models.

Similarly, AI is thoroughly entangled with global capitalism. Big data and machine learning are used in economic planning and decision-making, in customer, employee and logistics management, in lending and debt calculation and in financial markets, especially high frequency trading. Expectations for the future of these technologies are great, and there is no shortage of boosters promising tremendous, economic upheaval and growth (e.g., Schwab 2017). So thorough is their geographical diffusion and economic embeddedness that social theorists have begun thinking of them as infrastructure (Jaume-Palasi 2019). They underpin many new products and services, affect most aspects of private and professional life, and are entwined with our sense of progress and development (Elliott 2019).

The diverse and pervasive character of AI is closely related to their fragmented regulation; lack of oversight has allowed the technology to develop in this way, but the way that the technology has developed also makes oversight difficult to implement. As such, AI is regulated in different ways at different scales. Risks are anticipated by scientists and engineers during research and development as much as they are by social and legal structures in response to deployment. Design efforts to control AI include data filtering and cleaning, better and fairer algorithms, and the inclusion of a human operator in, on or off the process loop. At a regional or national scale, policy makers use subsidies, levies, and fines to attract or curtail technological test-bedding and implementation. In addition to government laws and policies, there are a range of less centralised efforts to regulate AI, for example, through ethical principles (e.g., the Asilomar Principles, and the OECD Principles on Artificial Intelligence), professional codes of conduct (e.g., Microsoft's The Future Computed, and SAP's Guiding Principles for Artificial Intelligence) and voluntary standards (e.g., 'ISO/IEC 23894 — Artificial Intelligence — Risk Management', and 'ISO/IEC 38507 — Governance implications of the use of artificial intelligence by organizations'). The demand for and effectiveness of these sites of regulation are further complicated by social norms, expectations, fears and desires. These interrelate in dynamic, complex ways, and it is important not to advance a too simplified model for their dynamics, in which ethics progress to standards which progress to law (cf. Winfield and Jirotka 2018), or regard them as a discrete patchwork of legal, social and ethical norms (cf. Gasser and Schmitt 2019).

Research on the regulation of AI risk is, like the regulation itself, in a formative stage. While it is clear that the risks posed are anticipated and mitigated in various ways, much remains unknown about how this occurs, what it means, and whether and how it is likely to change.

In this section, we have identified characteristics of AI that are relevant to a consideration of their social risks. These include their pervasiveness, economic embeddedness and currently fragmented regulation. More than this, however, we have also stressed that AI evade being known. The invisibilities of AI include their complexity, and inhuman and incomprehensible behaviour. These are important to acknowledge not only for what they reveal about our ignorance, but for how they inform decision-making about risk. In the next section, we return to the role of social knowledge in AI with these observations about their (in)visibilities in mind.

Renewing and pluralising the contribution of social science

The material characteristics of AI have implications for what can be known about them and the ways in which that knowing can be arrived at. In this section, we begin by reflecting on the epistemology of AI research and some of the limits to its knowledge production. We then identify a recent impasse between social science and AI industry. Rather than locate blame with

one side or the other, we use this as an opportunity to reassess the contribution of social science to AI research and regulation. We argue that taking ignorance into account offers one way forward.

It is very likely, given time, that areas of current ignorance about AI will yield to scientific enquiry. Further research on the technical and social infrastructures that AI are embedded within, for example, should help counteract unintended consequences and cascade failures. Similarly, work on tail risks and how AI fail should allow systems to be designed in more flexible and resilient ways. Three epistemic limits to the development of knowledge about AI need be acknowledged, however. The first is that not everything can be known about how an AI will behave before it is deployed. This is because of both practical reasons, related to time and cost, and methodological reasons, related to the incomprehensibility of trained neural networks. The second limit is that new knowledge is also likely to lead to a corresponding increase in non-knowledge (i.e., knowledge about our ignorance), as is typically the case in scientific research—the more we know, the more we are aware of our what we do not know. Finally, what lies beyond this knowledge and non-knowledge must always remain uncertain. There are unknown unknowns that are impossible to know (a domain of ignorance that Gross 2007 terms nescience). Some of these unknown unknowns may be discovered and known in retrospect, but because they cannot be anticipated their ramifications necessarily come as a surprise. From this it follows that some modesty about the conclusions of AI research is called for. Unfortunately, the very opposite appears to be occurring.

While social scientists have done much to reveal the shortcomings of technical risk mitigation, the relationship between social science and industry has cooled. For example, over the past five years, the annual state-of-the-sector reports published by the The AI Now Institute have become increasingly frustrated by the lack of improvement in corporate agendas, policies and workplace diversity (e.g., Crawford et al. 2019). More pointedly, in 2019, the philosopher Thomas Metzinger resigned from the European High-Level Expert Group on Artificial Intelligence, complaining of industry efforts to sideline and dilute objections made by academic participants. Questions have also been asked about why industry embraced the ethical turn in AI research in the first place. Powles and Nissenbaum (2018) have argued that the AI ethics boom has been used as a distraction from fundamental questions of social need. Kuziemski (2020) goes further, asserting that industry efforts have been part of a strategy to forestall direct government intervention. In keeping with these concerns, a call has been made for a second wave of algorithmic accountability to addresses the social context of problems in AI, and not only attempt to improve systems, but ask whether they should be built at all (Pasquale 2019).

These conflicts curb the possible contributions that social knowledge can make to AI research and regulation. On the one hand, it is likely that social scientists who tolerate industry interference will be ushered into less public positions from which their expertise can be formalised, quantified and co-opted to instrumental ends. On the other, those who resist industry may find themselves excluded from discussions, or elect to remain outside them, where they can better sharpen and direct their critique.

While involvement and antagonism both have their place, there are limits to what they can achieve. Social knowledge clearly has a role to play in the development of AI, just as interdisciplinary research is necessary in an ever more complex and fast-moving world. But to place one discipline at the service of another is to ignore the variety, complexity and spread of AI, and the range of expertise that is required to regulate and apprehend the risks that they pose. Moving social scientists into a subservient position may speed up development in the short-term. In the long term, however, it will undermine the fairness and safety of AI, and further jeopardise public trust in the sector. The co-optation of social knowledge to technical development reveals an immodest belief that all relevant social challenges have been identified and that it is only a matter of time before technical fixes are implemented.

Certain observations and arguments about the role of technology in society can only be made with critical distance from them. But the opposite is also true. With distance comes a lack of specificity and a different potential to affect change. Withdrawing from discussions with academic and industry developers may allow social scientists to be more faithful to their principles and politics, but it also obscures empirical details and removes dissenting voices from important discussions. While disengagement allows emphasis to be placed on key social issues and fundamental change, it is important not to discount the impact that internal voices can have in mediating and promoting social knowledge. In the short-term, the most effective way to minimise risks may be to adopt a more modest perspective on industry motivations and the social ramifications of AI, and contribute more actively to their development and regulation.

Rather than full co-optation or withdrawal, we suggest a different role, one that is both inside and outside, both committed and independent. For this role to be feasible would require greater acceptance of the ignorance of AI, not only from social scientists, but from everyone involved in efforts to make them fairer and safer. We proceed by thinking through how the invisibilities and visibilities of AI relate to knowledge and non-knowledge of the social.

Non-knowledge, understood as knowledge of ignorance for the purpose of improved decision-making, can help the invisibilities of AI to be better anticipated. The complexity and opacity of AI are dependent both on individual experience and expertise, and on how the edges of a given system are defined. Focusing on the code emphasises one set of skills, while focusing on the human, another. Similarly, the distribution of ignorance can be expected to vary depending on the system under examination. Complexity is thus understood as a relative and particular problematisation, which is better tackled through directed, interdisciplinary intervention, than general policies or solutions (e.g., widespread computer literacy). The inhuman and incomprehensible behaviour of AI suggests that more needs to be done to understand how AI fail in real-world settings. In this regard, we reinforce Mitchell's (2019) call for more research on tail risks, the brittleness of trained neural networks, the speed and severity of their failure, and how these are affected by such things as the size, accuracy and completeness of their training data, the frequency of their use, and the quality and quantity of their social interconnections. More than this, we argue that because unknown unknowns will always remain, the settings into which AI are deployed need to have proper measures in place to minimise or circumscribe the harm that their behaviour may cause.

The visibilities of AI, their pervasiveness, economic embeddedness and fragmented regulation, are all problems well suited to independent social knowledge. The great variation and widespread use of AI demands consideration of their current and future social consequences, without falling into the trap of assuming that this will be the same for all people and places. Because AI can cause considerable harm to individuals and groups, it is not sufficient to leave their development and regulation to those without expertise in this area. What is more, because AI systems are embedded within existing social, economic and political relations, we should be wary of ascribing too great of an importance to their technical deficiencies. It is certainly the case that some problems (e.g., harms of representation) evade technical solutions. But there are also problems which we would be well advised not to search for solutions for at all. Predictive policing, for example, is not only beset with data biases, but also contradicts what the social sciences tell us to be effective forms of policing (Asaro 2019). Rather than wasting resources making ineffective AI equally ineffective for all, social knowledge can help better integrate AI into the broader social policy landscape. Finally, the regulation of AI is not only something that social scientists should be contributing to, but is also a topic in need of further investigation. Regulation is currently fragmented and few efforts have been made to think reflexively and holistically about their various forms and scales, or how they interact with social worries and desires.

Better and more active recognition of the limits of AI, and the limits of what we know and can hope to know about them, could significantly improve their research and regulation. This

would require greater modesty from academia and industry as a way through the impasse that stands between them. There is also an opportunity for social knowledge and non-knowledge to help regulate some of the risky characteristics of AI. While we do not envisage that the potential contributions described here would eliminate the need for social scientific insight in the development of algorithms and models, nor mitigate the value of substantive critique, we do regard them as a healthy compliment that should help pluralise voices, mend divisions and minimise the harms that AI can cause. In our concluding remarks, we summarise the paper's argument and reflect on its political and ethical commitments.

Conclusion

The landscape of AI regulation is highly variegated and still in formation (Calo 2017; Daly et al. 2019). While many policy and governance recommendations have been made, in lieu of strong government regulation (that many worry will stifle innovation and competitiveness), no concrete set of mechanisms has emerged to curtail AI risks.

Social scientists have done much to raise awareness of the complex problems encountered and produced by AI outside the laboratory. This has involved a material and discursive push into the organisations, disciplines and debates through which these technologies are being developed. Substantial contributions have been made to characterise AI and propose measures for regulating the risks they generate, ranging from a programmed social contract to detailed data privacy legislation—such as the recent proposal made by the European Commission (EC. 2021). At the same time, their roll-out and adoption has been piecemeal, and little is known about their effects, both intended and otherwise, or how they are evolving and interacting.

There is both a lack of regulation, and considerable gaps in what we know and can hope to know about the risks posed by AI. Rather than be embarrassed by ignorance, we have maintained that it is a regular aspect of science and decision-making that should be acknowledged and incorporated into AI research and regulation. The recognition of ignorance responds to the invisibilities of AI, pushes back against bifurcations between social and technical knowledge, and resituates the role of the social scientist in AI research and regulation. The possibilities opened up by an exploration of both knowledge and non-knowledge of AI can be used to help to renew and pluralise the contributions that the social sciences are able to make.

Greater attention to ignorance is not a move without politics. Our use of social (non-)knowledge can be located, in the first instance, in an acceptance of the situatedness of knowledge and a greater humility with respect to its plural character. When technical progress is moving faster than regulation, immediate and certain answers are expected. To identify the contingent and embedded nature of knowledge may from this perspective be an unwelcome response. But, as stressed above, there is a need not only to rapidly solve technically defined risks, but also to ask wider questions about what is obscured by defining them in technical terms, providing wider understanding of their meanings and their possible implications. Being a multiscalar and transboundary phenomenon, crossing different areas of expertise and knowledge production, it is important not to foreclose any discipline's contribution to an understanding of what AI is and what it means for different actors.

To develop relevant and robust regulation, there is a need to not only reveal that AI, like many technologies, will have varied and often unjust distribution of benefits and risks, but also that the capacity and resources to govern AI differ within and between societies. It is also important to understand how different governing strategies are interpreted and acted upon by others are also important.

To develop relevant regulation of AI demands not only social knowledge about their pervasiveness, economic embeddedness and fragmented regulatory response, but a social non-knowledge that is attuned to their complexity, and inhuman and incomprehensible

behaviour. Allowing for our ignorance of their social implications, the regulation of AI can proceed in a more modest, situated, plural and ultimately robust manner.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

James M. White http://orcid.org/0000-0002-6735-3893
Rolf Lidskog http://orcid.org/0000-0001-6735-0011

References

Alcoff, Linda Martín. 2007. "Epistemologies of Ignorance: Three Types." In *Race and Epistemologies of Ignorance*, edited by Shannon Sullivan and Nancy Tuana, 39–57. SUNY Series, Philosophy and Race. Albany: State University of New York Press.

Ananny, Mike, and Kate Crawford. 2018. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20 (3): 973–989. doi:10.1177/1461444816676645.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asaro, Peter M. 2019. "AI Ethics in Predictive Policing: From Models of Threat to an Ethics of Care." *IEEE Technology and Society Magazine* 38 (2): 40–53. doi:10.1109/MTS.2019.2915154.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press. doi:10.1093/sf/soz162.

Boddington, Paula. 2017. *Towards a Code of Ethics for Artificial Intelligence. Artificial Intelligence: Foundations, Theory, and Algorithms*. Cham: Springer International Publishing.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, edited by Daniel D. Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugiyama, and Isabelle Guyon, 4349–57. New York: Curran Associates Inc.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the* World. Cambridge: The MIT Press.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1): 205395171562251. doi:10.1177/2053951715622512.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science (New York, N.Y.)* 356 (6334): 183–186. doi:10.1126/science.aal4230.

Calo, Ryan. 2017. "Artificial Intelligence Policy: A Primer and Roadmap." *U.C. Davis Law Review* 51: 399–435.

Collins, Harry. 2021. "The Science of Artificial Intelligence and Its Critics." *Interdisciplinary Science Reviews* 46 (1–2): 53–70. doi:10.1080/03080188.2020.1840821.

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv Preprint arXiv:1808.00023*.

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, et al. 2019. "AI Now 2019 Report." *AI Now Institute*. https://ainowinstitute.org/AI_Now_2019_Report.html.

Crawford, Kate. 2017. "The Trouble with Bias." In *Thirty-First Conference on Neural Information Processing Systems*, Long Beach, CA.

Daly, Angela, Thilo Hagendorff, Hui Li, Monique Mann, Vidushi Marda, Ben Wagner, Wei Wang, and Saskia Witteborn, Artificial Intelligence, Governance and Ethics: Global Perspectives. 2019. The Chinese University of Hong Kong Faculty of Law Research Paper No. 2019-15, University of Hong Kong Faculty of Law Research Paper No. 2019/033, Available at SSRN: https://ssrn.com/abstract=3414805 or http://dx.doi.org/10.2139/ssrn.3414805.

Dignum, Virginia. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms*. Cham: Springer International Publishing.

EC. 2021. *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence*. 3032/0106. Brussels, April 21. Accessed 25 May 2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Elish, Madeleine Clare. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5: 40–60. doi:10.17351/ests2019.260.

Elliott, Anthony. 2019. *The Culture of AI: Everyday Life and the Digital Revolution*. Oxon: Routledge.
Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.
Gasser, Urs, and Carolyn Schmitt. 2019. "The Role of Professional Norms in the Governance of Artificial Intelligence." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford University Press. Available at SSRN: https://ssrn.com/abstract=3378267 or http://dx.doi.org/10.2139/ssrn.3378267 .
Gross, Matthias, and Linsey McGoey. 2015. "Introduction." In *Routledge International Handbook of Ignorance Studies*, edited by Matthias Gross and Linsey McGoey, 1–14. Routledge International Handbooks. London: Routledge.
Gross, Matthias. 2007. "The Unknown in Process: Dynamic Connections of Ignorance, Non-Knowledge and Related Concepts." *Current Sociology* 55 (5): 742–759. doi:10.1177/0011392107079928.
JafariNaimi, Nassim. 2018. "Our Bodies in the Trolley's Path, or Why Self-Driving Cars Must *Not* Be Programmed to Kill." *Science, Technology, & Human Values* 43 (2): 302–323. doi:10.1177/0162243917718942.
Jaume-Palasi, Lorena. 2019. "Why We Are Failing to Understand the Societal Impact of Artificial Intelligence." *Social Research* 86 (2): 477–498.
Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage Publications.
Kitchin, Rob. 2017. "Thinking Critically about and Researching Algorithms." *Information, Communication & Society* 20 (1): 14–29. doi:10.1080/1369118X.2016.1154087.
Kröger, Wolfgang. 2021. "Automated Vehicle Driving: background and Deduction of Governance Needs." *Journal of Risk Research* 24 (1): 14–27. doi:10.1080/13669877.2020.1750465.
Kuziemski, Maciej. 2020. "The False Promise of 'Ethical AI'." *Project Syndicate*. https://www.project-syndicate.org/onpoint/false-promise-of-ethical-ai-by-maciej-kuziemski-2020-04.
Lanier, Jaron. 2020. "The Myth of AI: A Conversation with Jaron Lanier." *The Edge*. Accessed 2 August. https://www.edge.org/conversation/jaron_lanier-the-myth-of-ai.
Larkin, Brian. 2013. "The Politics and Poetics of Infrastructure." *Annual Review of Anthropology* 42 (1): 327–343. doi:10.1146/annurev-anthro-092412-155522.
Larsson, Stefan. 2017. *Conceptions in the Code: How Metaphors Explain Legal Challenges in Digital Times*. Oxford: Oxford University Press.
Lidskog, Rolf, and Göran Sundqvist. 2012. "Sociology of Risk." In *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, edited by Sabine Roeser, Rafaela Hillerbrand, Per Sandin, and Martin Peterson, 1001–1027. Dordrecht: Springer Netherlands.
Lindgren, Simon, and Jonny Holmström. 2020. "A Social Science Perspective on Artificial Intelligence: Building Blocks for a Research Agenda." *Journal of Digital Social Research* 2 (3): 1–15. doi:10.33621/jdsr.v2i3.65.
Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2): 205395171667967. doi:10.1177/2053951716679679.
Morgan, Phil, Chris Alford, and Graham Parkhurst. 2016. *Handover Issues in Autonomous Driving: A Literature Review*. Bristol: University of the West of England.
Neri, Hugo, and Fabio Cozman. 2019. "The Role of Experts in the Public Perception of Risk of Artificial Intelligence." *AI & Society* 35 (3): 663–673. doi:10.1007/s00146-019-00924-9.
Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,427–436.
Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
Orr, Will, and Jenny L. Davis. 2020. "Attributions of Ethical Responsibility by Artificial Intelligence Practitioners." *Information, Communication & Society* 23 (5): 719–735. doi:10.1080/1369118X.2020.1713842.
Pasquale, Frank. 2015. *The Black Box Society*. Cambridge: Harvard University Press.
Pasquale, Frank. 2019. "The Second Wave of Algorithmic Accountability." *Law and Political Economy*. https://lpeblog.org/2019/11/25/the-second-wave-of-algorithmic-accountability/.
Powles, Julia, and Helen Nissenbaum. 2018. "The Seductive Diversion of 'Solving' Bias in Artificial Intelligence." OneZero. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.
Rahwan, Iyad. 2018. "Society-in-the-Loop: Programming the Algorithmic Social Contract." *Ethics and Information Technology* 20 (1): 5–14. doi:10.1007/s10676-017-9430-8.
Reddy, Elizabeth, Baki Cakici, and Andrea Ballestero. 2019. "Beyond Mystery: Putting Algorithmic Accountability in Context." *Big Data & Society* 6 (1): 205395171982685. doi:10.1177/2053951719826856.
Renn, Otwin. 2021. "New Challenges for Risk Analysis: Systemic Risks." *Journal of Risk Research* 24 (1): 127–133. doi:10.1080/13669877.2020.1779787.
Russell, Stuart J., and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. 3rd ed. Harlow: Pearson Education.

Russell, Stuart. 2019. *Human Compatible: AI and the Problem of Control*. London: Penguin Books.
Scherer, Matthew U. 2016. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology* 29 (2): 353–400.
Schwab, Klaus. 2017. *The Fourth Industrial Revolution*. Geneva: World Economic Forum.
Schweizer, Pia-Johanna. 2021. "Systemic Risks – Concepts and Challenges for Risk Governance." *Journal of Risk Research* 24 (1): 78–93. doi:10.1080/13669877.2019.1687574.
Stilgoe, Jack. 2018. "Machine Learning, Social Learning and the Governance of Self-Driving Cars." *Social Studies of Science* 48 (1): 25–56. doi:10.1177/0306312717741687.
Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. "Intriguing Properties of Neural Networks." *arXiv:1312.6199*, February. http://arxiv.org/abs/1312.6199.
Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, et al. 2018. *AI Now 2018 Report*. AI Now Institute. https://ainowinstitute.org/AI_Now_2018_Report.html.
Winfield, Alan F. T., and Marina Jirotka. 2018. "Ethical Governance is Essential to Building Trust in Robotics and Artificial Intelligence Systems." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133): 20180085. doi:10.1098/rsta.2018.0085.
Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.