



Unsavory medicine for technological civilization: Introducing ‘Artificial Intelligence & its Discontents’

Shunryu Colin Garvey *

Stanford Institute for Human-Centered Artificial Intelligence

KEYWORDS Artificial Intelligence; history of science; Science & Technology Studies; modernity; well-being; medicine; critical theory; governance

This is, once again, the Age of Artificial Intelligence (Garvey 2018c). AI, the suite of techniques intended to make machines capable of performing tasks considered ‘intelligent’ when performed by people, is an epochal technology now colonizing an increasing number of domains, from Internet search and social media to the natural and social sciences; agriculture, banking, criminal sentencing, decision-making, and beyond. AI may soon become ubiquitous, coextensive with technological civilization itself: a taken-for-granted feature of modernity like running water or electricity.

But this does not mean that all is well. While AI promises to liberate and empower users, improve well-being, support social institutions, and enable sustainable development, it also threatens to automate and entrench precarity and illiberalism, degrade mental health, and accelerate the Earth’s ecological collapse.

Freud (1961) famously observed that civilization, despite being ostensibly intended to protect humanity from misery, is paradoxically a great source of unhappiness. Similarly, AI is both touted as the solution to humanity’s biggest problems and decried as one of the biggest problems humankind has ever faced – even, perhaps, its last. A plethora of pundits posit that AI poses an existential risk to human survival on this planet: If not nuclear war, climate catastrophe, or another global pandemic, then it will be ‘superintelligent’ machines that herald the Apocalypse (Barrat 2013; Bostrom 2014; Clark 2014; Yampolskiy 2015; Müller 2016; Cava 2018; Russell 2019). Or not. Other AI advocates claim a new wave of ethical developments will usher in the ‘Good AI Society’ (Floridi et al. 2018), free from scarcity and strife, thus bringing the West, as Japanese technologist Akihito Kodama (2016) has argued, to its teleological zenith: the return to Eden – abundance without work, life without pain – albeit digitized (Hilton 1964; Noble 1999; Geraci 2010; Diamandis and Kotler 2012).

CONTACT Shunryu Colin Garvey  shunryu@stanford.edu

*Guest Editor

© 2020 Institute of Materials, Minerals and Mining Published by Taylor & Francis on behalf of the Institute

Dystopian hellscape or Edenic utopia? Surely neither of these extremes are the only possible consequence of this vast, sociotechnical system of data, people, places, and things we call ‘AI’ – but what are the alternatives, and where are they to be found when expert partisans on each side dispute both the claims and the credentials of their counterparts? Who can help society make sense of the controversial technoscience of AI? The technoscientists whose careers depend on the success of AI? The business people who employ them? The policymakers devoted to the profits promised by the AI-powered ‘Fourth Industrial Revolution’ (World Economic Forum 2016; Schwab 2017; Mak n.d.; Brynjolfsson and McAfee 2012, 2016; McAfee and Brynjolfsson 2017; Brynjolfsson and Mitchell 2017; Ford 2009, 2015; though see also Wiener 1989; Gimpel 1977; Jenkins and Sherman 1979; White 1980; Johannesen 2019)?

None other than the critics. Little sense can be made of AI without reference to its *discontents* – those who doubt, question, challenge, reject, reform, and otherwise reprise ‘AI’ as it is practiced, promoted, and (re)produced. With the hope of scaffolding deeper, more nuanced understandings of both the epochal transformations being wrought by AI technologies and the range of responses required, possible, and as-yet unimagined, this special issue brings together critical accounts of AI and its discontents, past and present, in order to capture the significance of this historical moment, expand the horizons of the possible, and catalyze sociotechnical action on behalf of diverse publics and future generations whose autonomy – and humanity – are at stake.

AI criticism: the tradition of discontent

Often defined tongue-in-cheek by practitioners as ‘what computers can’t do, yet’ (Hendler and Mulvehill 2016), in its relentless focus on future horizons of technical capability, AI is amnesiac – scarcely aware of its own history, much less that of its critics. This may be a consequence of the fact that AI’s history has been written primarily by insiders and developers themselves (McCorduck 1979, 2004, 2019; Crevier 1993; Brooks 1999; Boden 1996; Nilsson 2010), all of whom tell triumphant stories of progress towards the current pinnacle upon which the world now stands, with a few bumps along the road thrown in for good measure.

According to this canon, AI began in 1956 at the Dartmouth Conference (Kline 2011), and there have only been two ‘bumps’ worth noting ever since – both, coincidentally, professors of philosophy at the University of California, Berkeley: Hubert Dreyfus (1965, 1972, 1992, 2007; Dreyfus and Dreyfus 1986, 1988), whose phenomenological attack on the Platonic formalism of early AI confounded its pioneers and presciently anticipated their failures; and John Searle (1980, 1999, 2014; Searle and Kurzweil 1999; Denton et al. 2002), whose ‘Chinese Room’ thought experiment ‘badly shook the little world of

[AI] by claiming and proving (so he said) that there was no such thing' (Motzkin and Searle 1989).

In fact, the tradition of AI criticism – to which the articles of this special issue make an important and timely contribution – is older, richer, and more diverse than suggested by the internalist history of AI (Hoos 1960a, 1960b, 1978; Wiener 1960; Greenberger 1962; Michael 1962; Bureau of Labor Statistics 1963; United States Congress Senate Committee on Labor and Public Welfare 1963; Neisser 1963; Ellul 1964; Terborgh 1965; Pierce et al. 1966; Silberman 1966; Hunt 1968; Jaki 1969; Wheeler 1972; Lighthill 1973; McDermott 1976; Glenn and Feldberg 1977; Noble 1978; Mori 1981; Ornstein, Smith, and Suchman 1984, 1985; Leontief and Duchin 1986; Bloomfield 1987; Born 1987; Suchman 1987; Beusmans and Wieckert 1989; Penrose 1989; Rosenbrock 1989; Brödner 1990; Corbett, Rasmussen, and Rauner 1991; Ennals 1991; Negrotti 1991; Collins 1992; Maturana and Varela 1992; Forsythe 1993; Ford 2001; Martin 1993; Bainbridge et al. 1994; Newquist 1994; Göranzon 1995; Göranzon and Florin 1990; 1991; Hutchins 1995; Edwards 1996; Hendriks-Jansen 1996; Kling 1996; Olazaran 1996; Adam 1998). Whereas Dreyfus and Searle were lambasted by AI partisans as ignorant outsiders (though see Armstrong, Sotala, and Ó hÉigeartaigh 2014), earlier critics, such as the mathematicians Richard Bellman (1958) and Hubert's brother Stuart (Dreyfus 1962, 2004, 2009, 2014), came from within the technical community. Perhaps the most fearsome of these discontents was Mortimer Taube (1911–1965), director of multiple research divisions of the Library of Congress, 'who, besides being an outstanding theorist and inventor, [was] one of the most successful business practitioners of the computer-based, data-processing art' (Solo 1963, 173; see also Smith 1993). Whereas Dreyfus, Searle, and many subsequent critics stayed squarely on theoretical terrain, Taube's now forgotten masterpiece, *Computers and Common Sense: The Myth of Thinking Machines* (1961), critiqued the social irresponsibility and economic profligacy of AI as well as its flawed philosophical foundations. Decades before 'deconstruction' became a term of art in academe, Taube analysed the AI literature to reveal its hidden assumptions, gaps in reasoning, and antiquated worldviews. Prior to Berger and Luckmann (1966), he showed how AI pioneers used language to socially construct the reality of 'thinking machines' through an interlocking web of peer-citation, long before theorists of the actor–network showed this to be a fundamental aspect of technoscientific power (Callon, Law, and Rip 1986). As it would become clear shortly after Taube's sudden death in 1965, the AI pioneers were defrauding everyone, from their military funders to their peers and the broader public, with overhyped claims about their machines. Yet Taube distinguished them from simple criminals by noting that while the creators of sophisticated literary and artistic forgeries are ostensibly aware that they are committing fraud, the creators of 'thinking machines' apparently believed they were actually doing science.

Dreyfus famously attacked the scientific credentials of AI by equating it to alchemy (1965). Like Darwin's Wallace, he arrived at this conclusion independently of Taube, who had argued several years prior that AI was a 'scientific aberration' like astrology or physiognomy (1961, 118–128). Taube looked forward to the day when its central dogma, the doctrine of 'Man-Machine Identity' – that the human brain is 'nothing but' a machine, and therefore can be simulated by another machine (76) – would not only be rejected, but, like the eugenics of biology or the colonial origins of anthropology, disowned entirely as an embarrassment to science. Understanding technoscience as a social activity that is ultimately meaningless if not helpful to society at large, Taube decried AI pioneers for using the term 'science' to 'peddle nostrums to a gullible public' while avoiding scrutiny by 'insisting on the pure scientific nature of their intentions' (124). To counteract their impairing influence on society, Taube attempted to introduce the criticism of technoscience 'as an enterprise similar in its aims to the established arts of literary, musical, art, and religious criticism,' one that 'views the [techno]scientific enterprise as an activity carried out by men [*sic*], not by demigods, nor even high priests' (v).

Widely influential at the time of its publication, the significance of Taube's critical project – which provides an excellent frame and sets a high bar for this special issue – was recognized by no less a technoscientist than Alvin M. Weinberg, Director of Oak Ridge National Laboratory, who observed that while the 'arts have always taken art critics and art criticism for granted,' technoscientists typically assume they have no need for critics:

Bad science is science that does not agree with nature; there are, in principle, objective criteria for deciding between good and bad science. But Taube's main contention is that in a field such as [AI] which deals with human artifacts (computers) and with logical, not empirical, issues, the tried-and-true criterion of agreement with experiment no longer serves to cull the bad from the good. Nor is the review of editors or fellow workers or government administrators sufficient—in Taube's opinion all are tainted with the same poison and, being taken in by the same alleged scientific fraud, can criticize only in detail, not in principle. If the scientific activities Taube criticizes were cheap, not much harm would be done; but since computers (like so much of modern Big Science) are expensive and are supported by public money, Taube argues that it is necessary and valid to subject these activities as a whole to the kind of criticism to which art is subjected, to criticize broadly the essential validity of the enterprise rather than to argue about the details within an accepted conceptual framework. That such a course is excruciatingly difficult, if for no other reason than that science is done by specialists and broad criticism of science must of necessity be done by people who know less than the specialists, does not deter Taube; he sees his duty and he states his opinions without pulling punches. (Weinberg 1962, 310)

Weinberg concluded by stating his hope that Taube's critique would go on to have influence far beyond the narrow confines of AI, for 'Much of modern Big Science could be helped by a dose of such unsavory, but necessary, medicine.'

Having edited and assembled the 10 articles in this special issue, I can assure the reader that each does its part to bring the bold critical enterprise begun by Taube into the twenty-first century – and that not a single one pulls its punches. Now allow me to echo Weinberg in hoping the contribution of these discontents proves broadly influential, for much of modern technoscience – and civilization itself – could be helped by the unsavory, but necessary, medicine offered herein.

Overview of the issue

In ‘The Lamp and the Lighthouse,’ Zachary Loeb examines the career of one of AI’s most notorious discontents, computer scientist Joseph Weizenbaum (1976), the first of a handful of notable defectors (e.g. Winograd and Flores 1987; Agre 1997; Suchman 2007; Marcus and Davis 2019; Smith 2019) from what he called the ‘Artificial Intelligentsia.’ Delving into Weizenbaum’s correspondence with the historian Lewis Mumford, whose own work affords considerable cultural context on the development of AI (Mumford 1963, 1964, 1965, 1967, 1970), Loeb provides an important corrective to the official canon: Weizenbaum was not – as so often portrayed by his former colleagues and adversaries in AI – a lone wolf, howling in the wilderness. Rather, his work, and the tradition of AI discontent more generally, was part of a larger tradition of social criticism responding to the accelerating automation, computerization, and complexification of technological civilization in the twentieth century.

Depending upon how ‘AI’ is defined, however, discontent is ancient (Wiener 1964; Cohen 1966; Winner 1989; Noble 1999; Herzfeld 2002; Geraci 2010; Russell and Norvig 2010; Garfinkel and Grunspan 2018). By excavating a minor literature on ‘Artificial Stupidity,’ Michael Falk’s article extends the critical tradition of the discontented beyond well-trod tomes like Forster’s ‘The Machine Stops’ (1909), the apocryphal ‘Book of the Machines’ in Butler’s *Erewhon* (1872), as well as Shelley’s 1818 *Frankenstein*, into the seventeenth and eighteenth centuries. Inverting narratives about the potential dangers of a hypothetical ‘superintelligence,’ Falk explicates the real risks of actual machine stupidity to open up a new (old) line of inquiry: Instead of always probing whether a given machine is truly intelligent – whatever that means – we ought instead to enquire, ‘What kind of stupid is it?’

Somewhat surprisingly, there is little if any consideration of stupidity in AI. But even more surprisingly, there is scarcely any more attention paid to what would seem to be the central concern of the entire AI enterprise: intelligence. Harry Collins, a sociologist and longtime discontent who critiques AI from the standpoint of Science and Technology Studies, summarizes and distills much of his oeuvre (1989, 1990, 1992, 2018; Collins and Kusch 1998) into ‘The Science of Artificial Intelligence and its Critics.’ It provides an important

tool for puncturing contemporary versions of the ‘myth of thinking machines’: a six-level scale of intelligence that clarifies what it actually means to be intelligent from a sociological perspective. In addition to combating hype by helping ordinary people distinguish fact from fiction in AI, Collins proposes that his framework could be used by experts to make AI a respectable science again – but do they have the ears to hear his ‘productive criticism’?

As Taube pointed out decades ago, hype-busting is important because the contemporary AI enterprise is not, and has never been, just a bit of harmless technoscientific experimentation (Garvey 2018a; Garvey and Maskal 2019). With its renewed geostrategic importance (Lee 2018; Scharre 2018; Comiter 2019; Garvey 2019b; Lin 2019; Mecklin 2019; NSCAI 2019; Prakash 2019; Bulletin of the Atomic Scientists 2020; Johnson 2020) and increasing presence in nearly every sector of society (Citron and Pasquale 2014; Pasquale 2015; Pasquinelli 2015; O’Neil 2016; Eubanks 2017; Tegmark 2017; Shoham et al. 2017, 2018; Wachter-Boettcher 2017; Broussard 2018; Foer 2018; S. U. Noble 2018; Taplin 2018; Susskind 2018; Vaidhyanathan 2018; Zuboff 2018; Atanasoski and Vora 2019; Benjamin 2019; Frey 2019; Perrault et al. 2019; Topol 2019; Webb 2019; Nourbakhsh and Keating 2020), enormous national budgets have been and are being planned on the promise that more AI will be ‘good’ for society (Webster et al. 2017; Dutton, Barron, and Boskovic 2018; The White House 2018; European Commission and Joint Research Centre 2018; Schmidt et al. 2020). Yet as Ulinicane and colleagues demonstrate in ‘Good Governance as a Response to Discontents? Déjà vu, or Lessons for AI from other Emerging Technologies,’ the ‘AI for Good’ narrative at the centre of these (inter)national initiatives is ahistorical, reproducing the amnesia of the AI canon by ignoring and resisting multiple relevant precedents in the governance of emerging technologies, such as public engagement and responsible innovation (Jasanoff 1996; Rowe and Frewer 2005; Wynne 2006; Stilgoe, Lock, and Wilsdon 2014; Özdemir and Springer 2018; Garvey 2019a).

Cheryl Holzmeyer’s ‘Beyond ‘AI for Social Good’ (AI4SG): Social Transformations – Not Tech-Fixes – for Health Equity,’ the fifth article in our special issue, similarly shows that these initiatives, however well-intentioned, breed discontent by distracting from root causes of social inequity and the meliorative potential of challenging existing systems of power. While advocates claim AI will improve public health, perhaps by making it possible to predict ‘well-being’ at the population level (Jaidka et al. 2020) or rapidly screen women for breast cancer (McKinney et al. 2020), all sociotechnical failures aside (Herper 2017; Ross 2018; Strickland 2019), the technical community’s narrow focus on potential downstream interventions contributes to neglect of the social determinants of health upstream, such as adequate income and housing, which in turn amplifies inequality and puts more of society at risk. In other words, although AI is supposed to be a powerful tool for improving well-being, by diverting attention and resources away from fundamentals

into expensive, experimental, expert-driven technical systems, it is paradoxically one of the greatest potential sources of social harm.

The next two articles expand upon the medical theme by exploring the role of AI in the clinic. In ‘Don’t Touch My Stuff: Historicizing Resistance to AI and Algorithmic Computer Technologies in Medicine,’ Ariane Hanemaayer documents the long tradition of doctors’ discontent as a struggle with and against machines over the authorship of medical truth. Rather than continuing to contest AI unaided, Hanemaayer suggests that physicians might find allies in their patients, with whom they share a partisan interest in reducing and protecting against the biases and other potential harms of medical AI systems (Coiera 1996; Cabitza, Rasoini, and Gensini 2017; Char, Shah, and Magnus 2018; Garvey 2018b; Krittanawong 2018). Saheli Datta Burton, Tara Mahfoud, and colleagues’ ‘Clinical Translation of Computational Brain Models: Understanding the Salience of Trust in Clinician-Researcher Relationships’ explores the physician’s predicament from another angle – the quintessential social bond of *trust* (Yamagishi 2011; Adali 2013). Drawing on their extensive experience with the Human Brain Project and interviews with experts, they show that without gaining clinicians’ trust through upstream collaboration that builds upon practitioners’ tacit knowledge, medical AI systems are unlikely to make meaningful contributions to patients’ health, even if they are adopted into the clinic.

The last three articles expand the frame of discontent to include the issues of biology, humanity, and identity. In ‘Truth from the Machine: Artificial Intelligence and the Materialisation of Identity,’ Keyes, Hitzig, and Blell explore how the natural sciences change when investigators utilize the quantitative techniques of AI to ‘discover’ qualitative social constructs such as ‘disease’ and ‘sexuality.’ Rachel Adams, in asking ‘Can Artificial Intelligence Be Decolonised?’, authoritatively unpacks the intertwined legacies of colonialism, racism, and Western cultural hegemony that underlie the conceptual foundations of AI, in order to establish an erudite theorization of what the strategy of decoloniality must mean for the field if it is not to be (re)appropriated as yet another ‘ethic.’ And finally, Alan Blackwell offers a personal account of his own discontent as a longtime practitioner that builds into a proposal for a future ethnography that would make it possible to think and conduct AI *otherwise*.

The future of AI and its discontents

Some discontents come from outside the field, others from within. Most if not all of them, however, critique AI in order to address larger issues of continued cultural relevance, such as the nature of ‘intelligence’; the development, implementation, and governance of large-scale sociotechnical systems (Garvey 2018d); the consequences of doing AI within the military-industrial-university complex; the problems of mind, brain, and consciousness in

a material universe; the relationship between language, thought, and society; as well as what it means to be ‘human’ in an increasingly computerized world.

These criticisms from the discontented reveal the deceptively simple two-letter moniker ‘AI’ to be a microcosm of technological civilization in dire need of strong medicine. While no single dose is strong enough to serve as an antidote to the cyclical malaise of modern machine-driven madness (Garvey 2018c), as AI grows more pervasive, its discontents will grow more numerous, and their critical prescriptions, however unsavory, ever more important to heed.

Acknowledgments

Shunryu Colin Garvey would like to thank the editor of ISR, Willard McCarty, for his patience, wisdom, and support on this special issue; Alan Blackwell for reading an early draft; Ariane Hanemaayer for her close collaboration and continued dedication to the project of ‘AI & its Discontents’; and all the contributors to this issue, as well as AI discontents past, present, and future.


Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Dr. *Shunryu Colin Garvey* is trained in the interdisciplinary field of Science & Technology Studies (STS). He uses AI as a case study to probe decision making under the conditions of complexity, uncertainty, and disagreement, in order understand how societies can more safely, fairly, and wisely govern controversial sociotechnical systems.

ORCID

Shunryu Colin Garvey  <http://orcid.org/0000-0002-7346-8873>

References

- Adali, Sibel. 2013. *Modeling Trust Context in Networks*. SpringerBriefs in Computer Science. New York, NY: Springer.
- Adam, Alison. 1998. *Artificial Knowing: Gender and the Thinking Machine*. New York, NY: Routledge.
- Agre, Philip. 1997. *Computation and Human Experience*. Learning in Doing: Social, Cognitive and Computational Perspectives. Cambridge: Cambridge University Press.
- Armstrong, Stuart, Kaj Sotala, and Seán S. Ó hÉigearthaigh. 2014. “The Errors, Insights and Lessons of Famous AI Predictions – and What They Mean for the Future.” *Journal of Experimental & Theoretical Artificial Intelligence* 26 (3): 317–342. <https://doi.org/10.1080/0952813X.2014.895105>.

- Atanasoski, Neda, and Kalindi Vora. 2019. *Surrogate Humanity: Race, Robots, and the Politics of Technological Futures*. Perverse Modernities. Durham, NC: Duke University Press.
- Bainbridge, William Sims, Edward E. Brent, Kathleen M. Carley, David R. Heise, Michael W. Macy, Barry Markovsky, and John Skvoretz. 1994. "Artificial Social Intelligence." *Annual Review of Sociology* 20 (1): 407–436. <https://doi.org/10.1146/annurev.so.20.080194.002203>.
- Barrat, James. 2013. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York, NY: Thomas Dunne Books.
- Bellman, Richard. 1958. "On 'Heuristic Problem Solving' by Simon and Newell." *Operations Research* 6 (3): 448–449.
- Benjamin, Ruha. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.
- Berger, Peter L., and Thomas Luckmann. 1966. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. 1st ed. Garden City, NY: Doubleday.
- Beusmans, Jack, and Karen Wieckert. 1989. "Computing, Research, and War: If Knowledge Is Power, Where Is Responsibility?" *Communications of the ACM* 32 (8): 939–951. <https://doi.org/10.1145/65971.65973>.
- Bloomfield, Brian P. 1987. *The Question of Artificial Intelligence: Philosophical and Sociological Perspectives*. New York, NY: Croom Helm.
- Boden, Margaret A., ed. 1996. *Artificial Intelligence*. 1st ed. San Diego, CA: Academic Press.
- Born, Rainer P., ed. 1987. *Artificial Intelligence: The Case Against*. New York, NY: St. Martin's Press.
- Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Brödnér, Peter. 1990. *The Shape of Future Technology: The Anthropocentric Alternative*. The Springer Series on Artificial Intelligence and Society. London: Springer London. <https://doi.org/10.1007/978-1-4471-1733-9>.
- Brooks, Rodney Allen. 1999. *Cambrian Intelligence: The Early History of the New AI*. Cambridge, MA: MIT Press.
- Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: The MIT Press.
- Brynjolfsson, Erik, and Andrew McAfee. 2012. *Race Against the Machine: How the Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy*. Lexington, MA: Digital Frontier Press.
- Brynjolfsson, Erik, and Andrew McAfee. 2016. *The Second Machine Age: Work, Progress, and Prosperity in the Time of Brilliant Technologies*. New York; London: W. W. Norton.
- Brynjolfsson, Erik, and Tom Mitchell. 2017. "What Can Machine Learning Do? Workforce Implications." *Science* 358 (6370): 1530–1534.
- Bulletin of the Atomic Scientists. 2020. "2020 Doomsday Clock Statement." Edited by John Mecklin. *Bulletin of the Atomic Scientists*, January, 20.
- Bureau of Labor Statistics. 1963. "Implications of Automation and Other Technological Developments: A Selected Annotated Bibliography." 1319–1. US Department of Labor.
- Butler, Samuel. 1872. *Erewhon*. London: Trübner.
- Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini. 2017. "Unintended Consequences of Machine Learning in Medicine." *JAMA* 318 (6): 517. <https://doi.org/10.1001/jama.2017.7797>.
- Callon, Michel, John Law, and Arie Rip, eds. 1986. *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*. London: The MacMillan Press Ltd.

- Cava, Marco Della. 2018. "Elon Musk Says AI Could Doom Human Civilization. Zuckerberg Disagrees. Who's Right?" *USA TODAY*, January 2, 2018. <https://www.usatoday.com/story/tech/news/2018/01/02/artificial-intelligence-end-world-overblown-fears/985813001/>.
- Char, Danton S., Nigam H. Shah, and David Magnus. 2018. "Implementing Machine Learning in Health Care – Addressing Ethical Challenges." *New England Journal of Medicine* 378 (11): 981–983. <http://doi.org/10.1056/NEJMp1714229>.
- Citron, Danielle Keats, and Frank A. Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89 (1): 1–33.
- Clark, Stuart. 2014. "Artificial Intelligence Could Spell End of Human Race – Stephen Hawking." *The Guardian*, December 2, sec. Science. <https://www.theguardian.com/science/2014/dec/02/stephen-hawking-intel-communication-system-astrophysicist-software-predictive-text-type>.
- Cohen, John. 1966. *Human Robots in Myth and Science*. London: George Allen & Unwin Ltd.
- Coiera, Enrico W. 1996. "Artificial Intelligence in Medicine: The Challenges Ahead." *Journal of the American Medical Informatics Association* 3 (6): 363–366. <https://doi.org/10.1136/jamia.1996.97084510>.
- Collins, Harry M. 1989. "Computers and the Sociology of Scientific Knowledge." *Social Studies of Science* 19 (4): 613–624.
- Collins, Harry M. 1990. *Artificial Experts: Social Knowledge and Intelligence Machines*. Cambridge, MA: The MIT Press.
- Collins, Harry M. 1992. "Hubert L. Dreyfus, Forms of Life, and a Simple Test for Machine Intelligence." *Social Studies of Science* 22 (4): 726–739.
- Collins, Harry M. 2018. *Artificial Intelligence: Against Humanity's Surrender to Computers*. Cambridge: Polity Press.
- Collins, Harry M., and Martin Kusch. 1998. *The Shape of Actions: What Humans and Machines Can Do*. Cambridge, MA: MIT Press.
- Collins, Randall. 1992. "Can Sociology Create an Artificial Intelligence?" In *Sociological Insight: An Introduction to Non-Obvious Sociology*, 155–188. New York, NY: Oxford University Press.
- Comiter, Marcus. 2019. *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It*. Cambridge, MA: Harvard Kennedy School, Belfer Center for Science and International Affairs.
- Corbett, J. Martin, Lauge Baungaard Rasmussen, and Felix Rauner. 1991. *Crossing the Border: The Social and Engineering Design of Computer Integrated Manufacturing Systems*. The Springer Series on Artificial Intelligence and Society. London; New York: Springer-Verlag.
- Crevier, Daniel. 1993. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York, NY: Basic Books.
- Denton, Michael, Thomas Ray, William A Dembski, John R Searle, George Gilder, and Ray Kurzweil. 2002. *Are We Spiritual Machines?: Ray Kurzweil vs the Critics of Strong A.I.* Edited by Jay W. Richards. 1st ed. Seattle, WA: Discovery Institute Press.
- Diamandis, Peter H., and Steven Kotler. 2012. *Abundance: The Future Is Better Than You Think*. 1st Free Press hardcover ed. New York, NY: Free Press.
- Dreyfus, Stuart. 1962. "Artificial Intelligence: Deus Ex Machina." *The Second Coming*, June.
- Dreyfus, Hubert L. 1965. *Alchemy and Artificial Intelligence*. P-3244. Santa Monica, CA: RAND Corporation.
- Dreyfus, Hubert L. 1972. *What Computers Can't Do: A Critique of Artificial Reason*. New York, NY: HarperCollins Publishers.

- Dreyfus, Hubert L. 1992. *What Computers Still Can't Do*. New York, NY: MIT Press.
- Dreyfus, Hubert L. 2007. "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian." *Philosophical Psychology* 20 (2): 247–268. <https://doi.org/10.1080/09515080701239510>.
- Dreyfus, Hubert L., and Stuart E. Dreyfus. 1986. "Competent Systems: The Only Future for Inference-Making Computers." *Future Generation Computer Systems* 2 (4): 233–243. [https://doi.org/10.1016/0167-739X\(86\)90023-3](https://doi.org/10.1016/0167-739X(86)90023-3).
- Dreyfus, Hubert L., and Stuart E. Dreyfus. 1988. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York, NY: Free Press.
- Dreyfus, Stuart E. 2004. "Totally Model-Free Learned Skillful Coping." *Bulletin of Science, Technology & Society* 24 (3): 182–187. <https://doi.org/10.1177/0270467604264813>.
- Dreyfus, Stuart E. 2009. "A Modern Perspective on Creative Cognition." *Bulletin of Science, Technology & Society* 29 (1): 3–8. <https://doi.org/10.1177/0270467608328708>.
- Dreyfus, Stuart E. 2014. "System 0: The Overlooked Explanation of Expert Intuition." In *Handbook of Research Methods on Intuition*, edited by Marta Sinclair, 15–27. Handbooks of Research Methods in Management. Cheltenham; Northampton, MA: Edward Elgar.
- Dutton, Tim, Brent Barron, and Gaga Boskovic. 2018. *Building an AI World: Report on National and Regional AI Strategies*. Toronto, ON: CIFAR. https://www.cifar.ca/docs/default-source/ai-society/buildinganaiworld_eng.pdf.
- Edwards, Paul N. 1996. *The Closed World: Computers and the Politics of Discourse in Cold War America*. Inside Technology. Cambridge, MA: MIT Press.
- Ellul, Jacques. 1964. *The Technological Society*. New York, NY: Vintage Books.
- Ennals, Richard. 1991. *Artificial Intelligence and Human Institutions*. The Springer Series on Artificial Intelligence and Society. London; New York, NY: Springer-Verlag.
- Eubanks, Virginia. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. 1st ed. New York, NY: St. Martin's Press.
- European Commission, and Joint Research Centre. 2018. "Artificial Intelligence: A European Perspective." http://publications.europa.eu/publication/manifestation_identifier/PUB_KJNA29425ENE.
- Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. "AI4People—an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28 (4): 689–707. <https://doi.org/10.1007/s11023-018-9482-5>.
- Foer, Franklin. 2018. *World Without Mind: The Existential Threat of Big Tech*. New York, NY: Penguin Books.
- Forsythe, Diana E. 2001. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, CA: Stanford University Press.
- Ford, Martin. 2009. *The Lights in the Tunnel: Automation, Accelerating Technology and the Economy of the Future*. Acculant Publishing.
- Ford, Martin. 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York, NY: Basic Books.
- Forster, E. M. 1909. "The Machine Stops." *The Oxford and Cambridge Review*, November.
- Forsythe, Diana E. 1993. "The Construction of Work in Artificial Intelligence." *Science, Technology & Human Values* 18 (4): 460–479.
- Freud, Sigmund. 1961. *Civilization and Its Discontents*. New York, NY: W. W. Norton.
- Frey, Carl Benedikt. 2019. *The Technology Trap: Capital, Labor, and Power in the Age of Automation*. Princeton, NJ: Princeton University Press.

- Garfinkel, Simson, and Rachel H Grunspan. 2018. *The Computer Book: From the Abacus to Artificial Intelligence, 250 Milestones in the History of Computer Science*. <https://www.overdrive.com/search?q=62FCFE29-6B04-40AC-BFA4-3FFCD565784B>.
- Garvey, Colin. 2018a. "AI Risk Mitigation Through Democratic Governance: Introducing the 7-Dimensional AI Risk Horizon." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 366–367. AIES '18. New York, NY: ACM. <https://doi.org/10.1145/3278721.3278801>.
- Garvey, Colin. 2018b. "Interview with Colin Garvey, Rensselaer Polytechnic Institute. Artificial Intelligence and Systems Medicine Convergence." *OMICS: A Journal of Integrative Biology* 22 (2): 130–132. <https://doi.org/10.1089/omi.2017.0218>.
- Garvey, Colin. 2018c. "Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996." *Technology's Stories*, March. <https://doi.org/10.15763/jou.ts.2018.03.16.18>.
- Garvey, Colin. 2018d. "A Framework for Evaluating Barriers to the Democratization of Artificial Intelligence." In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana USA - February 2-7, 2018*, 8079-8080. Palo Alto, CA: AAAI Press. <https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17320/16477>.
- Garvey, Colin. 2019a. "Hypothesis: Is 'Terminator Syndrome' a Barrier to Democratizing Artificial Intelligence and Public Engagement in Digital Health?" *OMICS: A Journal of Integrative Biology* 23 (7): 362–363. <https://doi.org/10.1089/omi.2019.0070>.
- Garvey, Colin. 2019b. "Artificial Intelligence and Japan's Fifth Generation: The Information Society, Neoliberalism, and Alternative Modernities." *Pacific Historical Review* 88 (4): 619–658. <https://doi.org/10.1525/phr.2019.88.4.619>.
- Garvey, Colin, and Chandler Maskal. 2019. "Sentiment Analysis of the News Media on Artificial Intelligence Does Not Support Claims of Negative Bias Against Artificial Intelligence." *OMICS: A Journal of Integrative Biology* 23: 1–16. <https://doi.org/10.1089/omi.2019.0078>.
- Geraci, Robert M. 2010. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. New York: Oxford University Press.
- Gimpel, Jean. 1977. *The Medieval Machine: The Industrial Revolution of the Middle Ages*. New York: Penguin Books.
- Glenn, Evelyn Nakano, and Roslyn L. Feldberg. 1977. "Degraded and Deskkilled: The Proletarianization of Clerical Work." *Social Problems* 25 (1): 52–64. <https://doi.org/10.2307/800467>.
- Göranzon, Bo, ed. 1995. *Skill, Technology, and Enlightenment: On Practical Philosophy*. The Springer Series on Artificial Intelligence and Society. London; New York: Springer-Verlag.
- Göranzon, Bo, and Magnus Florin, eds. 1990. *Artificial Intelligence, Culture and Language: On Education and Work*. The Springer Series on Artificial Intelligence and Society. London: Springer London. <https://doi.org/10.1007/978-1-4471-1729-2>.
- Göranzon, Bo, and Magnus Florin, eds. 1991. *Dialogue and Technology: Art and Knowledge*. The Springer Series on Artificial Intelligence and Society. London; New York: Springer-Verlag.
- Greenberger, Martin. 1962. *Computers and the World of the Future*. Cambridge, MA: MIT Press.
- Hendler, James, and Alice M. Mulvehill. 2016. *Social Machines: The Coming Collision of Artificial Intelligence, Social Networking, and Humanity*. Berkeley, CA: Apress.
- Hendriks-Jansen, Horst. 1996. *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. Complex Adaptive Systems. Cambridge, MA: MIT Press.

- Herper, Matthew. 2017. "MD Anderson Benches IBM Watson in Setback For Artificial Intelligence in Medicine." *Forbes*, February 19. <https://www.forbes.com/sites/matthewherper/2017/02/19/md-anderson-benches-ibm-watson-in-setback-for-artificial-intelligence-in-medicine/>.
- Herzfeld, Noreen L. 2002. *In Our Image: Artificial Intelligence and the Human Spirit*. Theology and the Sciences. Minneapolis, MN: Fortress Press.
- Hilton, Alice Mary. 1964. "Cyberculture – The Age of Abundance and Leisure." *Michigan Quarterly Review* 3 (4). <http://hdl.handle.net/2027/spo.act2080.0003.004:03>.
- Hoos, Ida R. 1960a. "The Impact of Office Automation on Workers." *International Labour Review* 82 (4): 363–388.
- Hoos, Ida R. 1960b. "The Sociological Impact of Automation in the Office." *Management Science* MT-1 (2): 10–19. <https://doi.org/10.1287/mantech.1.2.10>.
- Hoos, Ida R. 1978. "Paperwork Control." *Society* 16 (1): 5–8. <https://doi.org/10.1007/BF02712550>.
- Hunt, E. 1968. "Computer Simulation: Artificial Intelligence Studies and Their Relevance to Psychology." *Annual Review of Psychology* 19 (1): 135–168. <https://doi.org/10.1146/annurev.ps.19.020168.001031>.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.
- Jaidka, Kokil, Salvatore Giorgi, H. Andrew Schwartz, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2020. "Estimating Geographic Subjective Well-Being from Twitter: A Comparison of Dictionary and Data-Driven Language Methods." *Proceedings of the National Academy of Sciences* 117 (19): 10165–10171. <https://doi.org/10.1073/pnas.1906364117>.
- Jaki, Stanley L. 1969. *Brains, Minds, and Computers*. New York, NY: Herder and Herder, Inc.
- Jasanoff, Sheila. 1996. "Beyond Epistemology: Relativism and Engagement in the Politics of Science." *Social Studies of Science* 26 (2): 393–418.
- Jenkins, Clive, and Barrie Sherman. 1979. *The Collapse of Work*. London: Eyre Methuen.
- Johannessen, Jon-Arild. 2019. *The Workplace of the Future: The Fourth Industrial Revolution, the Precariat and the Death of Hierarchies*. Routledge Studies in the Economics of Innovation. Abingdon, Oxon; New York, NY: Routledge, Taylor & Francis Group.
- Johnson, James S. 2020. "Artificial Intelligence: A Threat to Strategic Stability." *Strategic Studies Quarterly* 24: 16–39.
- Kline, Ronald. 2011. "Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence." *IEEE Annals of the History of Computing* 33 (4): 5–16. <https://doi.org/10.1109/MAHC.2010.44>.
- Kling, Rob. 1996. *Computerization and Controversy: Value Conflicts and Social Choices*. San Diego, CA: Morgan Kaufmann.
- Kodama, Akihito. 2016. *Jinkō chinō wa watashitachi wo horobosunoka: keisanki ga kami ni naru 100 nen no monogatari* [Will AI Destroy us all? The 100 Year Story of Calculators Becoming Gods]. Tokyo: Diamond Publishers.
- Krittanawong, C. 2018. "The Rise of Artificial Intelligence and the Uncertain Future for Physicians." *European Journal of Internal Medicine* 48 (February): e13–e14. <https://doi.org/10.1016/j.ejim.2017.06.017>.
- Lee, Kai-Fu. 2018. *AI Superpowers: China, Silicon Valley, and the New World Order*. Boston, MA: Houghton Mifflin Harcourt.
- Leontief, Wassily, and Faye Duchin. 1986. *The Future Impact of Automation on Workers*. New York, NY: Oxford University Press.
- Lighthill, James. 1973. *Lighthill Report: Artificial Intelligence*. London: Science Research Council (SRC).

- Lin, Herbert. 2019. "The Existential Threat from Cyber-Enabled Information Warfare." *Bulletin of the Atomic Scientists* 75 (4): 187–196. <https://doi.org/10.1080/00963402.2019.1629574>.
- Mak, Alan. n.d. *Getting to the Future First: How Britain Can Lead the Fourth Industrial Revolution*. London: ConservativeHome: the home of conservatism.
- Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. 1st ed. New York, NY: Pantheon Books.
- Martin, C. Dianne. 1993. "The Myth of the Awesome Thinking Machine." *Communications of the ACM* 36 (4): 120–133. <https://doi.org/10.1145/255950.153587>.
- Maturana, Humberto R., and Francisco J. Varela. 1992. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Rev. ed. Boston, MA; New York, NY: Shambhala; Distributed in the U.S. by Random House.
- McAfee, Andrew, and Erik Brynjolfsson. 2017. *Machine, Platform, Crowd: Harnessing Our Digital Future*. New York, NY: W. W. Norton & Company.
- McCorduck, Pamela. 1979. *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. San Francisco, CA: W. H. Freeman.
- McCorduck, Pamela. 2004. *Machines Who Think: A Personal Inquiry Into the History and Prospects of Artificial Intelligence*. 25th anniversary update. Natick, MA: A.K. Peters.
- McCorduck, Pamela. 2019. *This Could Be Important: My Life and Times with the Artificial Intelligentsia*. Pittsburgh, PA: Carnegie Mellon University: ETC Press: Signature.
- McDermott, Drew. 1976. "Artificial Intelligence Meets Natural Stupidity." *ACM SIGART Bulletin* 57 (April): 4–9. <https://doi.org/10.1145/1045339.1045340>.
- McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, et al. 2020. "International Evaluation of an AI System for Breast Cancer Screening." *Nature* 577 (7788): 89–94. <https://doi.org/10.1038/s41586-019-1799-6>.
- Mecklin, John. 2019. "Dealing Realistically with the Artificial Intelligence Revolution." *Bulletin of the Atomic Scientists* 75 (3): 93–94. <https://doi.org/10.1080/00963402.2019.1604812>.
- Michael, Donald N. 1962. *Cybernation: The Silent Conquest*. Santa Barbara, CA: Center for the Study of Democratic Institutions.
- Mori, Masahiro. 1981. *The Buddha in the Robot*. 1st. English ed. Tokyo: Kosei Pub. Co.
- Motzkin, Elhanan, and John R. Searle. 1989. "Artificial Intelligence and the Chinese Room: An Exchange." *The New York Review of Books*, February 16. <http://www.nybooks.com/articles/1989/02/16/artificial-intelligence-and-the-chinese-room-an-ex/>.
- Müller, Vincent C. 2016. *Risks of Artificial Intelligence*. Boca Raton, FL: CRC Press.
- Mumford, Lewis. 1963. *Technics and Civilization*. 2nd ed. New York, NY: Harcourt, Brace & World, Inc.
- Mumford, Lewis. 1964. "Authoritarian and Democratic Technics." *Technology and Culture* 5 (1): 1–8.
- Mumford, Lewis. 1965. "Utopia, The City and The Machine." In *Utopias and Utopian Thought*, edited by Frank E. Manuel, 3–24. Boston, MA: Beacon Press.
- Mumford, Lewis. 1967. *Technics and Human Development*. Vol. 1. 2 vols. The Myth of the Machine. New York, NY: Harcourt, Brace & World.
- Mumford, Lewis. 1970. *The Pentagon of Power*. Vol. 2. 2 vols. The Myth of the Machine. New York, NY: Harcourt, Brace.
- Negrotti, Massimo, ed. 1991. *Understanding the Artificial: On the Future Shape of Artificial Intelligence*. The Springer Series on Artificial Intelligence and Society. London and New York: Springer-Verlag.
- Neisser, Ulric. 1963. "The Imitation of Man by Machine." *Science* 139 (3551): 193–197.

- Newquist, H. P. 1994. *The Brain Makers*. 1st ed. Indianapolis, IN: Sams Publishing.
- Nilsson, Nils J. 2010. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge; New York: Cambridge University Press.
- Noble, David F. 1978. "Social Choice in Machine Design: The Case of Automatically Controlled Machine Tools, and a Challenge for Labor." *Politics & Society* 8 (3-4): 313-347.
- Noble, David F. 1999. *The Religion of Technology: The Divinity of Man and the Spirit of Invention*. New York, NY: Penguin.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: New York University Press.
- Nourbakhsh, Illah Reza, and Jennifer Keating. 2020. *AI & Humanity*. Cambridge, MA: The MIT Press.
- NSCAI. 2019. *Interim Report*. Washington, DC: National Security Commission on Artificial Intelligence.
- Olazaran, Mikel. 1996. "A Sociological Study of the Official History of the Perceptrons Controversy." *Social Studies of Science* 26 (3): 611-659.
- O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Allen Lane, Penguin Books.
- Ornstein, Severo M., Brian C. Smith, and Lucy A. Suchman. 1984. "Strategic Computing." *Bulletin of the Atomic Scientists* 40 (10): 11-15. <https://doi.org/10.1080/00963402.1984.11459292>.
- Ornstein, Severo M., Brian C. Smith, and Lucy A. Suchman. 1985. "Strategic Computing: An Assessment." *Communications of the ACM* 28 (2): 134-136. <https://doi.org/10.1145/2786.314986>.
- Özdemir, Vural, and Simon Springer. 2018. "What Does 'Diversity' Mean for Public Engagement in Science? A New Metric for Innovation Ecosystem Diversity." *OMICS: A Journal of Integrative Biology* 22 (3): 184-189. <https://doi.org/10.1089/omi.2018.0002>.
- Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Pasquinelli, Matteo, ed. 2015. *Alleys of Your Mind: Augmented Intelligence and Its Traumas*. Lüneburg: Meson press.
- Penrose, Roger. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford; New York: Oxford University Press.
- Perrault, Raymond, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, and Juan Carlos Nieves. 2019. *AI Index 2019*. Stanford, CA: Stanford Institute for Human-Centered AI.
- Pierce, John R., John B. Carroll, Eric P. Hamp, David G. Hays, Charles F. Hockett, Anthony G. Oettinger, and Alan Perlis. 1966. *Language and Machines: Computers in Translation and Linguistics*. 1416. Washington, DC: Automatic Language Processing Advisory Committee, National Academy of Sciences, National Research Council.
- Prakash, Abishur. 2019. "The Geopolitics of Artificial Intelligence." *Scientific American Blog Network* (blog). July 11. <https://blogs.scientificamerican.com/observations/the-geopolitics-of-artificial-intelligence/>.
- Rosenbrock, H. H., ed. 1989. *Designing Human-Centred Technology: A Cross-Disciplinary Project in Computer-Aided Manufacturing*. The Springer Series on Artificial Intelligence and Society. London ; New York: Springer-Verlag.
- Ross, Casey. 2018. "IBM's Watson Recommended 'Unsafe and Incorrect' Cancer Treatments." *STAT*. July 25. <https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/>.

- Rowe, Gene, and Lynn J. Frewer. 2005. "A Typology of Public Engagement Mechanisms." *Science, Technology & Human Values* 30 (2): 251–290. <https://doi.org/10.1177/0162243904271724>.
- Russell, Stuart J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, NY: Viking.
- Russell, Stuart J., and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, NJ: Prentice Hall.
- Scharre, Paul. 2018. *Army of None: Autonomous Weapons and the Future of War*. 1st ed. New York ; London: W. W. Norton & Company.
- Schmidt, Eric, Robert O. Work, Safra Catz, Steve Chien, Mignon Clyburn, Christopher Darby, Kenneth Ford, et al. 2020. *NSCAI First Quarter Recommendations*. Washington, DC: National Security Commission on Artificial Intelligence.
- Schwab, Klaus. 2017. *The Fourth Industrial Revolution*. New York, NY: Crown Business.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–457.
- Searle, John R. 1999. "I Married a Computer." *The New York Review of Books*, April 8. <https://www.nybooks.com/articles/1999/04/08/i-married-a-computer/>.
- Searle, John R. 2014. "What Your Computer Can't Know." *New York Review of Books*, October.
- Searle, John R., and Ray Kurzweil. 1999. "I Married a Computer': An Exchange." *The New York Review of Books*, May 20. <http://www.nybooks.com/articles/1999/05/20/i-married-a-computer-an-exchange/>.
- Shoham, Yoav, Raymond Perrault, Erik Brynjolfsson, and Jack Clark. 2017. *AI Index 2017*. Stanford, CA: One Hundred Year Study on AI at Stanford University. <http://aiindex.org/2017/>.
- Shoham, Yoav, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz, and Zoe Bauer. 2018. *The AI Index 2018 Annual Report*. Stanford, CA: Stanford University.
- Silberman, Charles. 1966. *The Myths of Automation*. New York, NY: Harper & Row.
- Smith, Elizabeth S. 1993. "On the Shoulders of Giants: From Boole to Shannon to Taube: The Origins and Development of Computerized Information from the Mid-19th Century to the Present." *Information Technology and Libraries* 12 (2): 217–226.
- Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press.
- Solo, Robert A. 1963. "Automation: Technique, Mystique, Critique." *The Journal of Business* 36 (2): 166–178.
- Stilgoe, Jack, Simon J. Lock, and James Wilsdon. 2014. "Why Should We Promote Public Engagement with Science?" *Public Understanding of Science* 23 (1): 4–15. <https://doi.org/10.1177/0963662513518154>.
- Strickland, E. 2019. "IBM Watson, Heal Thyself: How IBM Overpromised and Underdelivered on AI Health Care." *IEEE Spectrum* 56 (4): 24–31. <https://doi.org/10.1109/MSPEC.2019.8678513>.
- Suchman, Lucy A. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge, UK: Cambridge University Press.
- Suchman, Lucy A. 2007. "Feminist STS and the Sciences of the Artificial." In *The Handbook of Science and Technology Studies*, edited by Edward J. Hackett, Olga Amsterdamska, Michael Lynch, and Judy Wajcman, 3rd ed., 139–163. Cambridge, MA: MIT Press: Published in cooperation with the Society for the Social Studies of Science.
- Susskind, Jamie. 2018. *Future Politics: Living Together in a World Transformed by Tech*. 1st ed. Oxford ; New York, NY: Oxford University Press.

- Taplin, Jonathan T. 2018. *Move Fast and Break Things: How Facebook, Google, and Amazon Cornered Culture and Undermined Democracy*. New York, NY: Little, Brown and Company.
- Taube, Mortimer. 1961. *Computers and Common Sense: The Myth of Thinking Machines*. New York, NY: Columbia University Press.
- Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. 1st ed. New York, NY: Alfred A. Knopf.
- Terborgh, George. 1965. *The Automation Hysteria*. New York, NY: W. W. Norton & Company, Inc.
- Topol, Eric J. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. 1st ed. New York, NY: Basic Books.
- United States Congress Senate Committee on Labor and Public Welfare. 1963. *Nation's Manpower Revolution: Hearings Before the Subcommittee on Employment and Manpower of the Committee on Labor and Public Welfare, United States Senate, Eighty-Eighth Congress, First Session, Relating to the Training and Utilization of the Manpower Sources of the Nation ...* U.S. Washington, DC: Government Printing Office.
- Vaidhyanathan, Siva. 2018. *Antisocial Media: How Facebook Disconnects US and Undermines Democracy*. New York, NY: Oxford University Press.
- Wachter-Boettcher, Sara. 2017. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. New York, NY: W. W. Norton & Company.
- Webb, Amy. 2019. *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity*. New York, NY: PublicAffairs.
- Webster, Graham, Rogier Creemers, Paul Triolo, and Elsa Kania. 2017. "Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)." *New America*. <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>.
- Weinberg, Alvin M. 1962. "Review: Computers and Common Sense: The Myth of Thinking Machines by Mortimer Taube." *The Library Quarterly* 32 (4): 309–310.
- Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco, CA: W.H. Freeman.
- Wheeler, Harvey. 1972. "Artificial Reasoning Machines and Politics." In *Computers and the Problems of Society*, edited by Harold Sackman, and Harold Borko, 467–493. Montvale, NJ: AFIPS Press.
- White, Lynn. 1980. *Medieval Technology and Social Change*. Reprinted. London: Oxford Univ. Press.
- The White House. 2018. "Artificial Intelligence for the American People." *The White House* (blog). May 10. <https://www.whitehouse.gov/briefings-statements/artificial-intelligence-american-people/>.
- Wiener, Norbert. 1960. "Some Moral and Technical Consequences of Automation." *Science* 131 (3410): 1355–1358.
- Wiener, Norbert. 1964. *God and Golem, Inc.: A Comment on Certain Points Where Cybernetics Impinges on Religion*. Cambridge, MA: MIT Press.
- Wiener, Norbert. 1989. *The Human Use of Human Beings: Cybernetics and Society*. London: Free Association.
- Winner, Langdon. 1989. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Seventh Printing. Cambridge, MA: MIT Press.
- Winograd, Terry, and Fernando Flores. 1987. *Understanding Computers and Cognition: A New Foundation for Design*. Reading, MA: Addison-Wesley.

- World Economic Forum. 2016. *The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution. Global Challenge Insight Report*. Geneva: World Economic Forum.
- Wynne, Brian. 2006. "Public Engagement as a Means of Restoring Public Trust in Science – Hitting the Notes, but Missing the Music?" *Community Genetics* 9 (3): 211–220. <https://doi.org/10.1159/000092659>.
- Yamagishi, Toshio. 2011. *Trust. The Science of the Mind*. Tokyo: Springer Tokyo. <http://link.springer.com/10.1007/978-4-431-53936-0>.
- Yampolskiy, Roman V. 2015. *Artificial Superintelligence: A Futuristic Approach*. Boca Raton, FL: CRC Press.
- Zuboff, Shoshana. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1st ed. New York, NY: PublicAffairs.



The lamp and the lighthouse: Joseph Weizenbaum, contextualizing the critic

Zachary Loeb

History and Sociology of Science, The University of Pennsylvania, Philadelphia, PA

ABSTRACT

The life and work of the computer scientist Joseph Weizenbaum is a testament to the ways in which the field of artificial intelligence has engendered discontent. In his articles, public talks, and most notably in his 1976 book, *Computer Power and Human Reason*, Weizenbaum challenged the faith in computerized solutions and argued that the proper question was not whether computers *could* do certain things, but if they *should*. As a computer scientist speaking out against computers, Weizenbaum has often been treated as something of a lone Jeremiah howling in the wilderness. However, as this article demonstrates, such a characterization fails to properly contextualize Weizenbaum. Drawing upon his correspondence with Lewis Mumford, this article argues that Weizenbaum needs to be understood as part of a community of criticism, while also exploring how he found the role of discontented critic to be a lonely and challenging one.

KEYWORDS

Joseph Weizenbaum; Lewis Mumford; artificial intelligence; the critique of technology; public intellectuals; technological pessimism; computers; ELIZA

A clear homage to Michelangelo's 'The Creation of Adam' dominated the first page of the 'Perspective' section of the November 13, 1977 edition of *The Baltimore Sun*, yet the hands that reached out towards one another did not belong to the naked Adam or to the bushy-bearded God. Instead, Adam had been replaced by a robot, while God had been replaced by a bald-bespectacled man in a lab coat. Beneath this provocative image, Joseph Weizenbaum sought to shine a light, for the *Sun*'s readers, on the hazards of being taken in by this high-tech retelling of 'let there be light' (Weizenbaum 1977b).

Between his role in composing the programming language SLIP, creating the influential proto-chatbot ELIZA, and teaching at MIT – Joseph Weizenbaum has secured an undeniable place in the history of computing, and artificial intelligence. Yet what most distinguishes Weizenbaum from other prominent figures in these fields is not his technical achievements, but the way he publicly turned against them (Crevier 1993; Kling 1996; McCorduck 2004).

Weizenbaum made his most lasting contribution as a critic, though what mattered to him was not the computer in and of itself but ‘the role of computers in our society’ (Weizenbaum and Wendt 2015, 33). It was precisely his intimate knowledge of the way that computers operated that led him to challenge the excited pronouncements that were made by those he referred to, with a mix of derision and respect, as the ‘Artificial Intelligentsia.’

Weizenbaum is one of AI’s most notable discontents. A rock in the field’s shoe which remains stubbornly present, he looms in the background of every dystopia-tinged article about AI, whispering, ‘I warned you.’ For those inclined to believe that computers and AI have set humanity on the wrong track, Weizenbaum appears as a hybrid of patron saint and Cassandra; while to those with high-hopes for computers and AI, who see ‘artificial intelligentsia’ as a badge of honour, Weizenbaum appears as a curmudgeonly Chicken Little. Drawing on Weizenbaum’s letters to his fellow social critic Lewis Mumford, this article argues that in order to properly understand Weizenbaum’s warnings it is necessary to recognize that he was not a lone prophet, howling in the wilderness, but a part of a broader milieu of social-criticism that saw science and technology as important sites for critique.

Lighting the lamp

In an anecdote about ELIZA which he returned to frequently, Weizenbaum recalled his secretary reacting as if he was interrupting a private conversation when he walked in on her using the program – despite the fact that she knew that the program did not understand the things she was saying to it (Weizenbaum and Wendt 2015, 90). Initially, Weizenbaum had expressed the belief that when the workings of a program were made clear ‘its magic crumbles away’ (1966, 36). However, in terms of the reactions to ELIZA, Weizenbaum was struck by how it ‘created the most remarkable illusion of having understood,’ which was ‘most tenaciously clung to among people who knew little or nothing about computers’ (1976, 189).

Named for the character from *Pygmalion*, ELIZA allowed a person to communicate in natural language with the computer and receive scripted responses. In its most famous version, ‘doctor,’ ELIZA took the form of a Rogerian psychotherapist – taking the statements of the user and echoing them back in the form of a question (Weizenbaum 1966). As far as Weizenbaum was concerned, the program ‘parodied’ a psychotherapist, and it disturbed him ‘that psychiatrists and other people took this parody seriously’ (Baumgartner and Payr 1995, 252).

Prior to ELIZA, Weizenbaum had constructed a program that excelled at playing the game Five-in-a-Row/Go-MOKU – his description of the program was published under the auspicious title ‘How to Make a Computer Appear Intelligent’ (1962). The key for this appearance boiled down to the

program's creator 'setting out to fool' users about the program's workings (24). Success was measured by how many people had been fooled and for how long (24). This program created 'a wonderful illusion of spontaneity,' and based on Weizenbaum's wry admission that he had trouble beating the program himself, the program seems to have played Five-in-a-Row fairly well (24). That a program should be able to fool a user into believing it to be intelligent, for Weizenbaum, was not proof of the program's intelligence but of the intelligence of the program's creator – though he still believed that the user would catch on eventually. This game-playing program helps underscore that what shocked Weizenbaum about ELIZA's reception was not that the program fooled its users, but *how willing those users were to be fooled*.

Granted, it is one thing to trick someone into thinking their opponent in a game is intelligent. Fooling someone into believing the therapist in whom they place their trust is intelligent is another. Though Weizenbaum's early admission that the point was to trick the user, his technical articles on ELIZA suggest he did not believe users would be consistently or willingly fooled. Sometimes 'magic crumbles' but sometimes the enchantment deepens, and ELIZA's powers were so entrancing that this sort of mystified misunderstanding has come to be known as 'the *Eliza* effect' (Broussard 2018; Murray 2017; Natale 2019; Suchman 2009; Turkle 2005; Wardrip-Fruin 2009). Rather than glorying in his success, Weizenbaum seemed chastened by it. ELIZA's only power was 'to deceive,' and he emphasized that 'no humane therapy of any kind ought to be grounded on' such a power (1977a, 354).

Though ELIZA quickened Weizenbaum's critical impulse, his concerns with the responsibilities of computer scientists predate the public reaction to the program. As a participant in a RAND Corporation Symposium in 1965, Weizenbaum was already giving voice to the concerns that make his reactions to ELIZA intelligible (McCracken 1965a, 1965b). Alluding to the doomsday clock of the Bulletin of the Atomic Scientists, Weizenbaum opined that it would only be a matter of time before 'computer people will start publishing a bulletin with the clock hands showing five minutes to twelve' (1965a, 24). This proximity to apocalyptic midnight was linked for Weizenbaum to his sense that computers were 'potentially dangerous instruments' though few others were willing to admit it (24). Over the course of the symposium, Weizenbaum echoed the belief, shared by the other speakers, that computers would become more widespread, but for him this was why it was incumbent upon computer professionals to learn to distinguish between 'what is worthwhile and what is not worthwhile,' 'what is right and wrong,' and to learn to say 'I will not do this' (25).

Referring to the doomsday clock, Weizenbaum noted 'Atomic scientists said after 1945 that they felt they had known sin' (24). It seems the ELIZA effect filled Weizenbaum with a similar feeling.

The lamp and the lighthouse

The concerns Weizenbaum raised regarding computers and AI transcend the moment in which they were originally posed, but his reference to the doomsday clock is a reminder that his thinking did not occur in a vacuum. The 1960s and 1970s were a period of social tumult, in the US and internationally, and it was a period in which many individuals in the technical and scientific community spoke out, even as a mixture of enthusiasm and mistrust greeted the latest technoscientific developments (Egan 2007; McCray and Kaiser 2016; Moore 2008; Roszak 1995; Schmalzer, Chard, and Botelho 2018; Tierney 2019; Turner 2008; Wisnioski 2016). As a critic from within the technical world, Weizenbaum has been seen as part of an intellectual tradition consisting of figures who have used their scientific status to bolster their criticisms of computers and AI – such as Norbert Wiener, Hubert Dreyfus, and some contemporary critics (Bynum 2001; Fleischman 2015; Frischmann and Selinger 2018; Gill 2019; Nadin 2019; Natale and Treré 2020; Mitcham 1994). While Wiener and Dreyfus certainly influenced him, to fully grapple with Weizenbaum's thinking it is necessary to place him in a broader milieu of twentieth-century social critics who were focused on technological issues.

From the outset of *Computer Power and Human Reason* (1976), Weizenbaum draws attention to the line of thinkers who have 'expressed grave concern about the conditions created by the unfettered march of science and technology.' This list included 'Mumford, Arendt, Ellul, Roszak, Comfort, and Boulding' (11), as well as Max Horkheimer and Erich Fromm. What unites this set of disparate thinkers is the attention they directed to the ways in which complex technological systems, and allegiance to technical thinking, resulted in the sacrifice of human needs on the altar of technological expedience. For many of these figures – several of whom, like Weizenbaum, were Jews who had fled from Nazi Germany – the Holocaust and the onset of the nuclear arms race stood as grim testaments to what occurred when the steely logic of mechanized rationality held power. While Weizenbaum clearly pays homage to these influences, his most significant comment about any of these figures is his expression of gratitude to 'Lewis Mumford, that grand old man' who Weizenbaum notes 'read all' of the manuscript (x). Weizenbaum gives Mumford further credit by noting that it was Mumford who had convinced him that 'it sometimes matters that a member of the scientific establishment say some things that humanists have been shouting for ages' (x).

Mumford was certainly one of the 'humanists' who had 'been shouting for ages.' By the time Weizenbaum set about writing *Computer Power*, Mumford had reached the twilight of his years as a public intellectual. Though Mumford was still publishing in the 1970s, his books that came out in that decade were largely collections of earlier work and autobiographical reminiscences (Mumford 1972, 1975, 1979, 1982). Throughout his oeuvre, few topics escaped

Mumford's critical gaze, he wrote about literature, art, architecture, politics, education, utopias, and much else – but he made his greatest impact with his writings about technology and cities (Blake 1990; Hughes and Hughes 1990; Jacoby 1987; Kuhns 1971; Luccarelli 1996; Miller 1989; Mitcham 1994; Wojtowicz 1996). Although he was aware that the approach would not win him many friends, Mumford tended to see himself in the mold of the prophet Jonah, someone 'who keeps uttering the very words you don't want to hear ... warning you that it will get even worse unless you yourself change your mind and alter your behavior' (Mumford 1979, 528).

Drawing on the thinking of his mentor Patrick Geddes, Mumford's early engagement with technology, saw him speaking of technological eras wherein a society could be defined by its sources of power and technological modes (Mumford 1934). Yet the onset of World War II and the rise of the nuclear era significantly darkened Mumford's worldview (Mumford 1939, 1940, 1946, 1951, 1954, 1956, 1959). Throughout the post-war period, Mumford worried about the power that was accumulating around complex technical systems, and his warnings about 'authoritarian technics' gradually turned into a theory of the 'Megamachine,' a vast social structure composed of 'human parts' that was remaking the world in its own mechanized image (Mumford 1964, 1967, 1970). A key feature of authoritarian technics was the danger that 'Once one opts for the system no further choice remains' (Mumford 1964, 6) – a development clearly visible in the growth of 'computer-dominated society' (Mumford 1970, 191). Indeed, for Mumford, the computer was the key device central to the functioning of the Megamachine (Loeb 2018). While he was concerned with specific machines, his greater concern was with the ideology that had developed around these machines: the belief that technoscience would solve all human problems. Mumford chafed at this 'myth of the machine,' which had become 'the basic religion of our present culture,' having 'so captured the modern mind that no human sacrifice seems too great provided it is offered up to the insolent Marduks and Molochs of science and technology' (Mumford 1975, 379). As a critic, Mumford confessed to having frequently felt rather lonely, yet in Weizenbaum he found a kindred spirit.

Friendships do not always leave behind many records, and it is not always possible for a historian to reconstruct the details of phone conversations or personal visits, but Weizenbaum's letters to Mumford provide a window into their relationship – despite the fact that Mumford's letters to Weizenbaum have not been preserved. The correspondence features many of the expected contents of a discourse between two friends, but it also includes ruminations on the state of the world that provide insight into Weizenbaum wrestling with his new identity as a social critic. Weizenbaum's letters make it clear that he was inspired by Mumford, praising him as 'an invaluable example to those others of us who

also struggle to become and remain free' even if that struggle should be painful (Weizenbaum to Mumford, 8/17/1972).¹

In much of their early correspondence, Weizenbaum provided Mumford with running commentary on the progress he was making on *Computer Power*. Though Weizenbaum had not been hiding his opinions on computers in the years prior to publication, he still expressed nervousness about what was to come, noting 'I am as frightened of the task ahead of me as layman might be were he called on to do surgery' (W2M, 11/21/1973). Yet Weizenbaum overcame his trepidation due to his certainty that what he was saying needed to be said, and this sentiment was bolstered by Mumford's belief that this message would have added weight coming from someone within the technical community. Weizenbaum found himself grappling with just how deeply entrenched the 'technological metaphor' had become amongst his technical peers; he worried that none believed 'reports of atrocities' until it was too late to prevent them, thus his task was to 'find a way of telling my readers that these particular atrocities are happening to them, that all this is not merely academic foolery' (W2M, 6/24/1974). The challenge of reaching the public is a leitmotif running throughout the correspondence – yet a definitive answer always eluded both men.

As Weizenbaum went about completing his book, the US was rocked by the resignation of President Nixon. While Weizenbaum was not upset to see Nixon go, his assessment of that situation featured the refusal to take good news at face value that was a hallmark of both correspondents' thinking. To Weizenbaum the risk of Nixon's resignation was 'that the people will think that all troubles are finally behind us,' whereas he saw nothing to celebrate in 'a system that elevates men like Mr. Nixon to the stewardship of the nation' (W2M, 8/8/1974). In a playfully apocalyptic comment, Weizenbaum pondered that 'What appears to be a dawn may be the herald of the coming Gotterdammerung!' (W2M, 8/8/1974). But in the midst of the world's tumults, with the war in Vietnam and the political fallout surrounding President Nixon looming in the background, working on *Computer Power* kept Weizenbaum grounded.

By the final months of 1974 Weizenbaum was circulating his manuscript to friends and students, he hoped the message of the book would resonate with readers both in and outside of the technical community as it had been composed with such a varied community of readers in mind. *Computer Power and Human Reason* is many things: an exploration of how Weizenbaum had come to criticize computing, a basic explanation of how computers work, a cutting commentary on the culture surrounding computing at prestigious universities, and most of all an impassioned plea to computer professionals to take responsibility for the things they were creating. Throughout the book, Weizenbaum upbraids his fellow computer professionals, not for thinking in terms of

¹Hereafter, correspondence from Weizenbaum to Mumford will be cited as (W2M, Date).

technical possibilities, but for failing to first ask whether it was appropriate to delegate certain tasks to machines; as he explained ‘since we do not now have any ways of making computers wise, we ought not now to give computers tasks that demand wisdom’ (1976, 227). And ‘wisdom’ was a trait that Weizenbaum doubted a computer could ever possess. A concern with responsibility animates the text, and Weizenbaum feared the ways in which ‘the myth of technological and political and social inevitability is a powerful tranquilizer of the conscience’ that serves ‘to remove responsibility from the shoulders of everyone who truly believes it’ (241). The book concludes with an appeal to readers, particularly those who are intimately involved with computers, to find the ‘simple kind of courage’ that involves learning to say ‘No!’ (276).

When he had started writing it, Weizenbaum had worried about how *Computer Power* would be received at MIT, as the book painted a less than flattering picture of those close to computers, and by arguing that they must assume responsibility the book suggested that they were currently failing to do just that. Yet, Weizenbaum seemed genuinely relieved and proud when he told Mumford how the graduate students studying AI with whom he had shared the book

have read it and not rejected its message out of hand. To the contrary, they all agree that the things it talks about ought at least to be on the agenda and that the book places them there in a responsible and thoughtful and thought provoking way. (W2M, 11/23/1974)

This response did not change Weizenbaum’s view of the looming, technoscientifically exacerbated risks of computing but he was heartened that the young people who had read the book seemed aware of those dangers. It suggested that his book could be a powerful corrective to the very tendencies he had critiqued.

Granted, it was not so much students’ reactions that Weizenbaum dreaded, but that of his colleagues. Weizenbaum humorously recounted an encounter with John McCarthy, ‘one of the founding cardinals of the artificial intelligence priesthood,’ who expressed a distaste for the book but who could not find any fault in it other than its style (W2M, 3/12/1976). According to Weizenbaum, McCarthy had accused him of learning his ‘writing style from Lewis Mumford,’ adding that ‘John was very puzzled that I seemed so pleased’ at this comment, which was not meant as a compliment (W2M, 3/12/1976). Yet McCarthy’s comment on style, and his allusion to Mumford, speaks not so much to a ‘fault’ as a faultline: the growing gulf between Weizenbaum and ‘the artificial intelligence priesthood.’ Whereas the former had become a proud heretic, the latter group was only becoming more fervent in its AI evangelism.

Though Weizenbaum expressed amusement at McCarthy comparing his work to Mumford’s, what buoyed him was not this idle comparison to his

idol, but a host of other experiences. Weizenbaum noted that since the book's publication he had spoken with many colleagues who had thanked him for his work, telling him that they shared many of his concerns, but 'have not uttered them publicly' (W2M, 3/12/1976). It seemed to Weizenbaum, based on these conversations, that his book was for many of these unnamed individuals 'a kind of legitimation of their own ideas and even of having nonconventional ideas at all' (W2M, 3/12/1976). Though Weizenbaum expressed to Mumford that he was pleased to take on this role, he also noted how saddened he was to see that so many of his colleagues were unwilling to speak out against the accepted orthodoxy. Their response further demonstrated 'how urgent it is for everyone who can do so to set examples – and how enormously powerful examples can be' (W2M, 3/12/1976).

While attending the Rose-Hulman Institute of Technology's conference on 'Technology at the Turning Point,' Weizenbaum found himself once more reminded of the reasons why many of his colleagues, even if they agreed with him, refrained from uttering their views publicly. In this regard as well, he recognized that Mumford had paved the path he was now treading. As he described it to Mumford, Weizenbaum delivered a talk attacking the idea of technological inevitability and warning that 'popular attitudes toward computers have converted them to instruments which tell us what we must hope for ... Having been told what we must hope for, we immediately turn computers into instruments which realize our hopes' (W2M, 4/6/1976). Yet in recounting the event to Mumford, Weizenbaum did not dwell on his own comments, but instead expressed frustration at the talk delivered by the historian Melvin Kranzberg, who had 'railed against the enemies of technology,' amongst whom Mumford was identified 'most emphatically' (W2M, 4/6/1976). This was certainly not the first time that Kranzberg had criticized Mumford publicly. Several years earlier, Kranzberg had called those who opine on technology's impact on human values 'intellectual Luddites,' framing them as 'concerned, articulate, but ineffectual' (Kranzberg 1964, 580). That he placed Mumford within this group is made clear in his review of the first volume of Mumford's *Myth of the Machine*, where Kranzberg fondly recalled Mumford's earlier work – before calling the current Mumford 'a prophet of doom' and asking 'Has something gone wrong with technology? Or just the prophet?' (Kranzberg 1967, 687).

While Kranzberg's earlier critiques of Mumford had been printed, the conference provided Weizenbaum an opportunity to rise to Mumford's defense, challenge Kranzberg's reading of Mumford's work, as well as his technological optimism. For Weizenbaum, Kranzberg's framing of critics as 'enemies of technology' served as an easy way for the celebrants of technology to avoid having to engage with the merits of those critiques. For Weizenbaum the accusation of 'technological determinist,' which Kranzberg hurled at Mumford, overlooked Mumford's lifelong commitment to searching for alternative technological

modes, while also overlooking how certain technological systems are created in order to remake society in their own image. Commenting on Kranzberg's rhetorical techniques, Weizenbaum described them as a 'very neat trick' that frames 'every good that technology has brought about as a product of the essence of technology, while any evil that might be mentioned was naturally a consequence of the weakness and the fallibility of man' (W2M, 4/6/1976). Beyond being a tactic to dismiss criticism, Weizenbaum felt that assessments such as Kranzberg's permitted 'the technological optimist to charge ahead without thought to the consequences his magic may engender' (W2M, 4/6/1976). The encounter was a discomfiting reminder that views such as his and Mumford's were still quite unwelcome within the technoscientific community.

Despite some cheering comments from students, words of support from colleagues, and some positive reviews of the book, encounters of the sort that Weizenbaum had with McCarthy and Kranzberg had a draining impact. Whereas Mumford had ample experience sounding the siren in vain, the failure of *Computer Power* to overcome the cultural influence of the artificial intelligentsia and their ilk left Weizenbaum deeply depressed. Reflecting on Kranzberg's comments, Weizenbaum told Mumford that he had not been 'particularly upset' by them as 'you and I both know that there is an enormous amount of this kind of nonsense being spoken every day' (W2M, 4/6/1976). Considering the body of letters Weizenbaum sent to Mumford, it is hard to accept his claim that he was mostly unbothered. While noting that he was doing his best 'to guard against becoming a "true believer" and a victim of my own skepticism,' Weizenbaum nevertheless found the 'superficiality' with which most people approached computers compelled him to darken his vision (W2M, 10/2/1977). Many of Weizenbaum's public writings feature their fair share of glum assessments, but few equal the mournful heft with which he told Mumford, 'I have the insuppressible impression that the world is getting exponentially worse while my efforts to save an individual life here and there and a modicum of sanity generally can be increased only linearly' (W2M, 3/18/1978) – a combinatorial explosion of technoscientific development his critique could not overcome in real time.

In his initial letters to Mumford, Weizenbaum's words were tinged with optimism. Though he had appreciated the enormity of the task before him, and the risks to his reputation it entailed, as Weizenbaum dispatched draft after draft to Mumford he was animated by a faith that speaking out would make a difference. In the aftermath of the book's publication, however, the flames of hope diminished into frail sparks, especially as the years marched on. Weizenbaum continued to derive succour from Mumford's work and their friendship. Yet it seemed that another area in which Weizenbaum came to emulate Mumford was in feeling like the mocked prophet who watches in solitude as the very things they glumly warned of come to pass. Weizenbaum feared that

‘the world is rushing madly and apparently thoughtlessly into a whole series of crises, into all of them with increasing acceleration,’ and though he hoped that he could contribute ‘to the rise of the general level of sanity in our society’ he was increasingly doubtful it would be enough (W2M, 1/19/1984). Looking about the world Weizenbaum asked ‘If one has only ten fingers, which holes in the dam which holds back the tide of barbarism should one attempt to plug?’ (W2M, 1/19/1984).

Where once Weizenbaum had asked Mumford for guidance on how to be a public intellectual, he later sought guidance on how to be at peace with being ignored. Weizenbaum and Mumford both tried to plug the holes created by an overly eager embrace of science and technology, but even their best attempts had not stopped the rising tide from pouring through.

Illuminating metaphors

At the conclusion of ‘The Last Dream,’ an article in which Weizenbaum couched the history of AI in humanity’s long obsession with creating new life, he delivered a punchy sentence that neatly summarizes his outlook: ‘Real life is not computable’ (2019, 193). Throughout the article, a copy of which he sent to Mumford, Weizenbaum emphasized that technoscience was not the only way to understand human experience, and, moreover, that the attempt shamefully cheapened that experience. He critiqued prominent members of the Artificial Intelligentsia for overestimating the capabilities of AI, but noted that they were correct in recognizing ‘the readiness of people to believe such things, both then and now’ (181).

Weizenbaum’s understanding of the proclivity to put faith in the pronouncements of the Artificial Intelligentsia is one of many reasons his thinking remains important today. Yet what truly makes his perspective on this significant is not his comments on the matter, but the ways in which he explained how this had come to pass. Technological enthusiasm courses through much of American history (Adas 2006; Hughes 1989; Marvin 1988; Segal 2005), and though this optimistic/utopian streak has periodically been challenged by a contrary perspective (Ezrahi, Mendelsohn, and Segal 1994; Kling 1996; Winner 1989) – Weizenbaum’s forlorn letters to Mumford capture how the critics felt their warnings could not overcome the widespread enthusiasm for technoscience. In addition to the Mumford-focused confrontation with Kranzberg, Weizenbaum had run-ins with his own ideological opponents as well. In some cases these played out in the pages of popular periodicals (Feigenbaum and McCorduck 1983; Weizenbaum 1983a, 1983b) and soured the relationship in perpetuity (McCorduck 2019; Weizenbaum and Wendt 2015). And in still other cases Weizenbaum’s viewpoints seemed to be included just so they could be shot down. For example, his article ‘Once more – a computer revolution’ (1978), reprinted in the volume *The Computer Age* (Dertouzos and Moses

1979), was followed by two responses that disagreed with him, whereas no other chapter in the book was followed by similar retorts.

Curmudgeonly though Weizenbaum could be, he was no misanthrope. There is a genuine note of disbelief in his comment to Mumford that 'it's hard to understand sometimes how the world can be so screwed up in spite of the fact that most people are good' (W2M, 3/25/1980). This matter is doubtlessly one which has concerned many individuals, yet, at least when it came to technology, Weizenbaum and Mumford had an answer. In Mumford's estimation 'the ultimate religion of our seemingly rational age' was 'the myth of the machine,' wherein an idolatrous belief in the power of science and technology to solve all problems resulted in 'progress' becoming a shorthand for technical achievements alone (1970). Over the two volumes of the aptly titled *Myth of the Machine*, Mumford had gone into detail explaining how this myth was maintained by a new priesthood that had enshrined the computer as its new god (1967, 1970). Explaining how this grim situation could come into being, Mumford described how this new system 'bribed' the masses by offering them a share in the technologically produced bounty in exchange for their complicity (1970). Far from rejecting Mumford's analysis, Weizenbaum deepened it, drawing explicitly on the idea of the myth of the machine to argue that 'technological metaphors ... and technique itself ... so thoroughly pervade our thought process that we have finally abdicated to technology the very duty to formulate questions' (1972, 611). The challenge of formulating questions was thus rendered moot as the technological system only permitted questions to which it was itself the answer. As Weizenbaum pithily put it, 'the computer has almost since its beginning been basically a solution looking for a problem,' thus it framed every problem as one that only a computer could solve and presented itself as the solution (ben-Aaron 1985, 2). Just as someone with a hammer sees every problem as a nail, to the Artificial Intelligence, every problem looks computable.

Weizenbaum stressed the importance of investigating who the victims and victors 'of our much-advertised technological progress' would be; how the computer and AI would impact our idea of ourselves; what 'irreversible forces' would come into existence thanks to 'our worship of high technology'; the sort of world these decisions would leave our descendants; and the importance of asking 'what *limits* ought' to be imposed on the application of computation to 'human affairs' (1978, 19) – precisely because such questions do not give themselves to easily computable answers. No technological determinist, Weizenbaum emphasized that much was dependent on the choices being made by technologists and he sought to remind his peers that 'since the beginning of recorded history, decisions having the most evil consequences are often made in the service of some overriding good' (1976, 273). Weizenbaum, much like Mumford, underscored that the technological situation in which we find ourselves is not the result of inevitability, or the result of irresistible

progress, but the result of *decisions*. And different decisions could have been, and still can be, made.

There are certainly aspects of Weizenbaum's thought that have aged oddly, which is why it is important to see him not as a lone Jeremiah howling in the desert but as a member of a community of critics. Weizenbaum's frequent allusions to the Nazis and the Holocaust can appear anachronistic to those who forget that Weizenbaum fled Nazi Germany in his youth, and his statement 'we have learned nothing. Civilization is as imperiled today as it was then' (1976, 256) may appear hyperbolic to those who view the Nazis as an aberration instead of as a product of modernity. Nevertheless, the technological critiques by numerous Jewish critics in the post-war period drew clear links between the experience of the Holocaust and a distrust that new technoscientific advances could be handled responsibly (Anders 2010, 2013; Arendt 1958; Fromm 1968; Jonas 1985). Zygmunt Bauman, for example, drew directly on Weizenbaum to warn of the dehumanizing risks of computers (Bauman 1989, 115–116). While Weizenbaum saw himself as leaning on the work of a range of social critics, his own writing entered into that conversation, and eventually many of those who Weizenbaum had drawn on wound up citing him in return (Ellul 1990; Roszak 1994).

Nevertheless, Weizenbaum's work remains vital today, not because he was an early critic of AI, but because the questions he raised concerning AI and computerized society are still the ones that trouble technological civilization. Computers remain powerful metaphors for describing our world (Chun 2011; Golumbia 2009; Hong 2020; Mosco 2005). Hype around artificial intelligence continues to overstate the capabilities of these systems (Broussard 2018). The myths around AI distract from the humans doing the work (Gillespie 2018; Roberts 2019). The complexity of various systems continues to deepen and exacerbate social divisions (Noble 2018; Pasquale 2015). All the while, computers and AI remain alluring devices onto which societies project their hopes for social transformation (Benjamin 2019; Ames 2019).

Where once Weizenbaum cited past critics in order to place himself in a continuum of technoscientific critique, many contemporary critics cite Weizenbaum to similarly situate themselves (Loeb 2015, 28). Yet what keeps Weizenbaum relevant today is not only his place amongst past critics, but that his critiques still speak directly to the social implications computers and AI. It is an enduring relevance that Sun-ha Hong captures by evoking Weizenbaum's 'counsel' that 'not everything that *can* be done with technology ought to be done' (Hong 2020, 186). Contemplating the 'paradoxical role of computers,' Weizenbaum observed

on the one hand the computer makes it possible in principle to live in a world of plenty for everyone, on the other hand we are well on the way to using it to create a world of suffering and chaos. (1983c, 10)

Nearly 40 years after those words were written, many of those studying computers and AI are still wrestling with this paradox.

For Weizenbaum, much like Mumford, technologies could never be considered in isolation from the societies in which they were found. Thus, it was impossible to assess computers or AI merely by looking at machines. As Weizenbaum put it ‘The computer ... is a mirror in which certain aspects and qualities of contemporary America are reflected’ he went on to add that his purpose was not ‘to say anything good or bad about the mirror. I am talking about what it reflects’ (Mulvihill 1986, 135). As for what it reflected? Weizenbaum minced no words, noting that ‘the computer is embedded in our crazy society ... and this society is obviously insane’ (Weizenbaum and Wendt 2015, 42).

A light in the dark

In describing the computer metaphor Weizenbaum warned how it had become a ‘lamppost under whose light, and only under whose light, men seek answers to burning questions’ (1976, 158). Considering the darkness threatening to envelop the world, it was understandable that people would flock to the glimmer of that lamppost and satisfy themselves with whatever illumination it could bring. Yet, Weizenbaum was not content to ‘seek answers’ under that glow.

If only certain answers could be cast into the brightness of the lamppost of the computer metaphor, then it was necessary to find other light sources. Thus, in explaining what he hoped to accomplish in his work, Weizenbaum told Mumford that ‘all I have done is to carry a lamp into a dark fortress’ (W2M, 9/2/1974). Noting how Mumford had expressed the belief that merely carrying that lantern is ‘an important thing to do,’ he also credited Mumford with being one of ‘the keepers of its light’ (W2M, 9/2/1974). In expressing gratitude to Mumford for his guidance and assistance, Weizenbaum noted that they had been brought together by a shared set of thoughts and worries. ‘I came to know yours as a lost sailor comes to know a lighthouse on whose wide ranging beam he has come to rely long before he ever [became] its keeper’ (W2M, 11/23/1974). The lamp that Weizenbaum held aloft burnt brightly for a time, and he had sought not only to rally his fellow computer scientists but the broader public as well. Thus, beneath the image of the scientist and the robotic Adam, Weizenbaum had sought to illuminate for the readers of *The Baltimore Sun* that ‘Perhaps we have even more to fear from people who act as if they were computers than from computers programmed to pretend to be human’ (Weizenbaum 1977b, K2). Despite his efforts to hold the lantern high, the darkness grew as more and more people came to gather around the lamppost of the computer metaphor. And though Weizenbaum strived to be a keeper of the light, he confided that ‘I lean toward the view these days that the lights are going out’ (W2M, 1/19/1984).

Whether one agrees or disagrees Weizenbaum's assessment, his work remains a lantern worth rekindling as we make our way through this dark fortress. This lantern may not tell us the correct path to take, but it will at the very least remind us that we do not have to stay on our present path. That so much of what awaits us remains uncertain and unpredictable only heightens the need for us to carefully evaluate the route we shall take. And thus the question remains, who will be willing to tread this lonely path, littered with accusations of alarmism, in order to hoist this lantern anew?

Let there be light.

It is sorely needed.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Zachary Loeb is a PhD Candidate in the History and Sociology of Science Department of The University of Pennsylvania. His research sits at the intersection of the history of technology and disaster studies, with focus on critics of technology. He is working on a dissertation project on the year 2000 computing crisis (Y2K).

References

Archival Material

Lewis Mumford Collection, University of Pennsylvania, Charles Patterson Van Pelt Library, Department of Special Collections, Philadelphia.

Books and Articles

- Adas, M. 2006. *Dominance by Design: Technological Imperatives and America's Civilizing Mission*. Cambridge: Belknap Press of Harvard University Press.
- Ames, M. 2019. *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*. Cambridge: MIT Press.
- Anders, G. 2010. *Die Antiquiertheit des Menschen 1: Über die Seele im Zeitalter der zweiten industriellen Revolution*. München: Verlag C.H. Beck.
- Anders, G. 2013. *Die Antiquiertheit des Menschen 2: Über die Zerstörung des Lebens im Zeitalter der dritten industriellen Revolution*. München: Verlag C.H. Beck.
- Arendt, H. 1958. *The Human Condition*. Chicago, IL: University of Chicago Press.
- Bauman, Z. 1989. *Modernity and the Holocaust*. Ithaca, NY: Cornell University Press.
- Baumgartner, P., and S. Payr, eds. 1995. *Speaking Minds: Interviews with Twenty Eminent Cognitive Scientists*. Princeton, NJ: Princeton University Press.
- ben-Aaron, D. 1985. "Weizenbaum Examines Computers and Society: Interview." *The Tech*, April 9, 2.
- Benjamin, R. 2019. *Race After Technology*. London: Polity Press.
- Blake, C. N. 1990. *Beloved Community: The Cultural Criticism of Randolph Bourne, Van Wyck Brooks, Waldo Frank, and Lewis Mumford*. Chapel Hill: University of North Carolina Press.

- Broussard, M. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge: MIT Press.
- Bynum, T. W. 2001. "Computer Ethics: Its Birth and Its Future." *Ethics and Information Technology* 3: 109–112.
- Chun, W. H. K. 2011. *Programmed Visions: Software and Memory*. Cambridge: MIT Press.
- Crevier, D. 1993. *AI: The Tumultuous History of the Search for Artificial Intelligence*. New York: Basic Books.
- Dertouzos, M., and J. Moses, eds. 1979. *The Computer Age: A Twenty-Year View*. Cambridge: MIT Press.
- Egan, M. 2007. *Barry Commoner and the Science of Survival: The Remaking of American Environmentalism*. Cambridge: MIT Press.
- Ellul, J. 1990. *The Technological Bluff*. Grand Rapids, MI: William B. Eerdmans.
- Ezrahi, Y., E. Mendelsohn, and H. Segal, eds. 1994. *Technology, Pessimism, and Postmodernism*. Amherst: University of Massachusetts Press.
- Feigenbaum, E., and P. McCorduck. 1983. "Computers in Your Future." *The New York Review of Books*, December 8.
- Fleischman, W. 2015. "Just Say 'No!' to Lethal Autonomous Robotic Weapons." *Journal of Information, Communication and Ethics in Society* 13 (3/4): 299–313.
- Frischmann, B., and E. Selinger. 2018. *Re-Engineering Humanity*. Cambridge: Cambridge University Press.
- Fromm, E. 1968. *The Revolution of Hope: Toward a Humanized Technology*. New York: Harper & Row.
- Gill, K. 2019. "From Judgment to Calculation: The Phenomenology of Embodied Skill." *AI and Society* 34: 165–175.
- Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.
- Golumbia, D. 2009. *The Cultural Logic of Computation*. Cambridge: Harvard University Press.
- Hong, S. 2020. *Technologies of Speculation: The Limits of Knowledge in a Data-Driven Society*. New York: New York University Press.
- Hughes, T. 1989. *American Genesis: A Century of Invention and Technological Enthusiasm*. New York: Penguin Books.
- Hughes, T., and A. Hughes. 1990. *Lewis Mumford: Public Intellectual*. New York: Oxford University Press.
- Jacoby, R. 1987. *The Last Intellectuals: American Culture in the Age of Academe*. New York: Basic Books.
- Jonas, H. 1985. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago, IL: University of Chicago Press.
- Kling, R. 1996. *Computerization and Controversy: Value Conflicts and Social Choices*. 2nd ed. San Diego: Academic Press, Inc.
- Kranzberg, M. 1964. "Technology and Human Values." *The Virginia Quarterly Review* 40 (4): 578–592.
- Kranzberg, M. 1967. "Review: Man and Megamachine." *The Virginia Quarterly Review* 43 (4): 686–693.
- Kuhns, W. 1971. *The Post-Industrial Prophets: Interpretations of Technology*. New York: Weybright and Talley.
- Loeb, Z. 2015. "Introduction." In *Islands in the Cyberstream: Seeking Havens of Reason in a Programmed Society*, edited by J. Weizenbaum and G. Wendt. Duluth: Litwin Books. <https://www.boundary2.org/2018/07/loeb/>.
- Loeb, Z. 2018. "From Megatechnic Bribe to Megatechnic Blackmail: Mumford's 'Megamachine' After the Digital Turn." *Boundary 2 (Online)*. The Digital Turn, July 30.

- Luccarelli, M. 1996. *Lewis Mumford and the Ecological Region: The Politics of Planning*. New York: Guilford Press.
- Marvin, C. 1988. *When Old Technologies Were New: Thinking About Electric Communication in the Late Nineteenth Century*. Oxford: Oxford University Press.
- McCorduck, P. 2004. *Machines Who Think*. Boca Raton, FL: CRC Press.
- McCorduck, P. 2019. *This Could Be Important: My Life and Times with the Artificial Intelligentsia*. Pittsburgh: Carnegie Mellon University: ETC Press.
- McCracken, D. 1965a. "The Rand Symposium: Defining the Problem." *Datamation* 8 (8): 24–30.
- McCracken, D. 1965b. "The Rand Symposium: Part Two." *Datamation* 8 (9): 66–73.
- McCray, P., and D. Kaiser. 2016. *Groovy Science: Knowledge, Innovation and American Counterculture*. Chicago, IL: University of Chicago Press.
- Miller, D. 1989. *Lewis Mumford: A Life*. New York: Weidenfeld and Nicolson.
- Mitcham, C. 1994. *Thinking Through Technology: The Path between Engineering and Philosophy*. Chicago, IL: University of Chicago Press.
- Moore, K. 2008. *Disrupting Science: Social Movements, American Scientists, and the Politics of the Military, 1945–1975*. Princeton, NJ: Princeton University Press.
- Mosco, V. 2005. *The Digital Sublime: Myth, Power, and Cyberspace*. Cambridge: MIT Press.
- Mulvihill, R., ed. 1986. *Reflections on America, 1984: An Orwell Symposium*. Athens: University of Georgia Press.
- Mumford, L. 1934. *Technics and Civilization*. New York: Harcourt, Brace.
- Mumford, L. 1939. *Men Must Act*. New York: Harcourt, Brace.
- Mumford, L. 1940. *Faith for Living*. New York: Harcourt, Brace.
- Mumford, L. 1946. *Values for Survival*. New York: Harcourt, Brace.
- Mumford, L. 1951. *The Conduct of Life*. New York: Harcourt, Brace.
- Mumford, L. 1954. *In the Name of Sanity*. New York: Harcourt, Brace.
- Mumford, L. 1956. *The Transformations of Man*. New York: Harper and Brothers.
- Mumford, L. 1959. "An Appraisal of Lewis Mumford's *Technics and Civilization* (1934)." *Daedalus* 88 (3): 527–536.
- Mumford, L. 1964. "Authoritarian and Democratic Technics." *Technology and Culture* 5 (1): 1–8.
- Mumford, L. 1967. *Technics and Human Development*. Vol. 1 of *The Myth of the Machine. Technics and Human Development*. New York: Harvest/Harcourt Brace Jovanovich.
- Mumford, L. 1970. *The Pentagon of Power*. Vol. 2 of *The Myth of the Machine. Technics and Human Development*. New York: Harvest/Harcourt Brace Jovanovich.
- Mumford, L. 1972. *Interpretations and Forecasts: 1922–1972*. New York: Harcourt, Brace and Jovanovich.
- Mumford, L. 1975. *Findings and Keepings: Analects for an Autobiography*. New York: Harcourt, Brace and Jovanovich.
- Mumford, L. 1979. *My Work and Days: A Personal Chronicle*. New York: Harcourt, Brace, Jovanovich.
- Mumford, L. 1982. *Sketches from Life: The Autobiography of Lewis Mumford*. New York: Dial Press.
- Murray, J. 2017. *Hamlet on the Holodeck: The Future of Narrative in Cyberspace*. Cambridge: MIT Press.
- Nadin, M. 2019. "Machine Intelligence: A Chimera." *AI and Society* 34: 215–242.
- Natale, S. 2019. "If Software is Narrative: Joseph Weizenbaum, Artificial Intelligence and the Biographies of ELIZA." *New Media and Society* 21 (3): 712–728.
- Natale, S., and S. Treré. 2020. "Vinyl Won't Save us: Reframing Disconnection as Engagement." *Media, Culture and Society* 42: 1–8.

- Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.
- Roberts, S. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.
- Roszak, T. 1994. *The Cult of Information: A Neo-Luddite Treatise on High-Tech, Artificial Intelligence, and the True Art of Thinking*. Berkeley: University of California Press.
- Roszak, T. 1995. *The Making of a Counterculture*. Berkeley: University of California Press.
- Schmalzer, S., D. Chard, and A. Botelho, eds. 2018. *Science for the People: Documents from America's Movement of Radical Scientists*. Amherst, MA: University of Massachusetts Press.
- Segal, H. 2005. *Technological Utopianism in American Culture*. Syracuse, NY: Syracuse University Press.
- Suchman, L. 2009. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge: Cambridge University Press.
- Tierney, M. 2019. *Dismantlings: Words Against Technology in the American Long Seventies*. Ithaca, NY: Cornell University Press.
- Turkle, S. 2005. *The Second Self: Computers and the Human Spirit*. Cambridge: MIT Press.
- Turner, F. 2008. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago, IL: University of Chicago Press.
- Wardrip-Fruin, N. 2009. *Expressive Programming: Digital Fictions, Computer Games, and Software Studies*. Cambridge: MIT Press.
- Weizenbaum, J. 1962. "How to Make a Computer Appear Intelligent." *Datamation* 8 (2): 24–26.
- Weizenbaum, J. 1966. "ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9 (1): 36–45.
- Weizenbaum, J. 1972. "On the Impact of the Computer on Society." *Science* 176: 609–614.
- Weizenbaum, J. 1976. *Computer Power and Human Reason: From Judgement to Calculation*. San Francisco, CA: W.H. Freeman.
- Weizenbaum, J. 1977a. "Computers as 'Therapists'." *Science, New Series* 198 (4315): 354.
- Weizenbaum, J. 1977b. "In Search of the True Thinking Machine." *The Baltimore Sun*. November 13.
- Weizenbaum, J. 1978. "Once More—A Computer Revolution." *Bulletin of the Atomic Scientists*, September.
- Weizenbaum, J. 1983a. "Computers in Your Future." *The New York Review of Books*, December 8.
- Weizenbaum, J. 1983b. "The Computer in Your Future." *The New York Review of Books*, October 27.
- Weizenbaum, J. 1983c. "The Paradoxical Role of the Computer." Holst Memorial Lecture 1983. Technische Hogeschool Eindhoven, December 14.
- Weizenbaum, J. 2019. "The Last Dream." *AI and Society* 34: 177–194.
- Weizenbaum, J., and G. Wendt. 2015. *Islands in the Cyberstream: Seeking Havens of Reason in a Programmed Society*. Duluth: Litwin Books.
- Winner, L. 1989. *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought*. Cambridge: MIT Press.
- Wisnioski, M. 2016. *Engineers for Change: Competing Visions of Technology in 1960s America*. Cambridge: MIT Press.
- Wojtowicz, R. 1996. *Lewis Mumford and American Modernism: Eutopian Themes for Architecture and Urban Planning*. Cambridge: Cambridge University Press.



Artificial stupidity

Michael Falk

School of English, University of Kent, Canterbury, UK

ABSTRACT

Public debate about AI is dominated by Frankenstein Syndrome, the fear that AI will become superhuman and escape human control. Though superintelligence is theoretically possible, it distracts from a more pressing problem: the rise of Artificial Stupidity (AS). This article discusses the cultural roots of Frankenstein Syndrome, and provides a conceptual framework for evaluating the stupidity of artificial agents. It then identifies an alternative literary tradition that exposes the perils and benefits of AS. In the writings of Edmund Spenser, Jonathan Swift and E.T.A. Hoffmann, ASs *replace, enslave or delude* their human users. More optimistically, Joseph Furphy and Laurence Sterne imagine ASs that can augment human intelligence by serving as *maps* or as *pipes*. These writers provide a strong counternarrative to the myths that currently drive the AI debate. They identify ways even stupid agents can thwart human aims, and demonstrate the social and scientific value of literary texts.

KEYWORDS

Artificial intelligence; stupidity; English literature; German literature; Australian literature; superintelligence; singularity; cognitive artefacts

Frankenstein Syndrome

To misquote Joseph Weizenbaum: what could it mean to speak of stupidity when one speaks of machines? The question is rarely asked. It is intelligent machines that capture the imagination, not stupid ones – machines that interpret X-rays, generate fake photographs, trade stocks, or win at Chess, Go and *Starcraft II*. Journalists and intellectuals tell stories of such machines to suggest that AI is possible, probably imminent, and potentially far more intelligent than humanity. According to Max Tegmark, for instance, DeepMind's AlphaGo system is proof that machines have already achieved genuine 'intuition' and 'creativity' (2018, 87). In 2016, AlphaGo defeated Go grandmaster Lee Sedol 4-1, using risky moves that 'def[ied] millennia of human intuition' (2018, 87–88).¹ On the flipside, when people *do* tell stories of machine stupidity, their purpose is normally just to refute the AI-believers. In 2017, for instance, *The Economist* published work by an AI to demonstrate that humans still write better copy (The Economist 2017). Both Tegmark and *The Economist* address

CONTACT Michael Falk  m.g.falk@kent.ac.uk  School of English, Rutherford College, University of Kent, Canterbury, Kent CT2 7NX, UK

¹For a more balanced assessment of AlphaGo's intellect, see Mitchell (2019, 214–18).

© 2020 Institute of Materials, Minerals and Mining Published by Taylor & Francis on behalf of the Institute

the same concern: When will the machines outsmart us? Neither seem overly interested in a somewhat more pressing problem: If the machines haven't outsmarted us yet, then in what way are they stupider?

The AI debate is warped by Frankenstein Syndrome, or the fearful fascination with superintelligent agents. Over the last two decades, a string of bestselling authors have predicted the arrival of superintelligent AI (Kurzweil 2006; Bostrom 2014; Tegmark 2018; Russell 2019), to the thunderous applause of celebrity businessmen and intellectuals like Bill Gates, Elon Musk, Stephen Hawking and Sam Harris. These prophets of superintelligence often claim they are being ignored (Bostrom 2014, v; Russell 2019, 132–44), but in reality their fears dominate the public imagination. One measure of their dominance comes from the cinema, where for 10 years Marvel's superhero films have commanded the global box office ('List of Highest-Grossing Franchises and Film Series' 2020). In these films, superintelligent AIs such as Ultron, Jarvis/Vision, Arnim Zola, the Supreme Intelligence and the mysterious 'algorithm' from *Captain America: Winter Soldier* (2014) continually threaten humanity and indeed the universe. Frankenstein Syndrome is a problem because it draws attention away from a more pressing concern. Though superintelligent AI may be possible in theory, Artificial Stupidity (AS) already exists, is continually infiltrating new corners of society, and is still only poorly understood.

The Syndrome is rooted in an old and persistent cultural myth. There are long traditions of writing about 'automata', or self-moving machines, stretching back to ancient China, India, Greece and Israel (Kang 2011; Mayor 2018), but two centuries ago, these traditions took off in a new direction with the publication of Mary Shelley's *Frankenstein* (Shelley 1998). Frankenstein's monster was a new kind of automaton, for two reasons:

- (1) He was rooted in modern science, in particular the new sciences of 'chemistry' and 'electricity' (Shelley 1998, 32, 24). These new sciences had exposed natural forces that were strange and fluid enough to conceivably explain consciousness, and yet were also controllable enough to drive the real technological advances of the Industrial Revolution.
- (2) He was endowed with conscious intelligence, with 'reason', 'sensations', 'perceptions' and 'passions' (Shelley 1998, 79, 114, 119). In fact his intelligence is superhuman. In only few months, he is able to progress from absolute ignorance – 'I knew, and could distinguish, nothing' (Shelley 1998, 80) – to a high level of literacy and cunning. He is so cunning, in fact, that even the most intelligent human in the story – Victor Frankenstein – is powerless to thwart him.

Frankenstein's monster can be described as the first modern superintelligence, an electrical supermind coursing over a chemical substrate. His arrival fundamentally altered the terms of the 'control problem'. Since ancient times,

writers had considered the risk that automata might escape human control (Kang 2011, 21; Mayor 2018, 29–30, 206), but here for the first time was an automaton whose intelligence would make control impossible, and who could conceivably be manufactured in the near future by a scientific process.² This frightening being quickly became a powerful myth.

We need a cure for Frankenstein Syndrome. While the fear of Frankenstein's monster has dominated the discussion, a different kind of artificial agent has steadily been colonizing every aspect of human life: autopilots that keep planes on course, but rob human pilots of their skill (Fry 2019, 155–57); infuriatingly useful autocorrect; sputtering automated faucets and stingy towel dispensers; intrusive and occasionally frightening targeted advertisements; insipid home assistants like Siri and Alexa. These artificial agents are regularly billed as AIs due to their ability to respond flexibly to their environment, but as I will show, their apparent intelligence is also a kind of Artificial Stupidity (AS).

Not only has AS received far less attention than superintelligence, stupidity itself is a neglected topic: 'Basic points about stupidity's place in the conceptual field [remain] unclear' (Golob 2019, 564). How is stupidity related to ignorance, cleverness, forgetfulness, intelligence or imagination? As it turns out, stupidity has long been a preoccupation of novelists and poets, and literature holds a rich store of ideas about what it means for a human or a machine to be stupid.

Frankenstein Syndrome is a literary disease, and a literary disease requires a literary cure. In what follows, I unearth an alternative tradition of literary works that explore the perils and potentials of AS. In the first section, I define AS. I introduce the distinction between stupidity of understanding and stupidity of judgement, and show in what way modern AS can be said to suffer from stupidity of judgement. In the next section, I consider three literary examples of AS run wild: Edmund Spenser's *The Faerie Queene* (Spenser 1977), Jonathan Swift's *Gulliver's Travels* (Swift 2005), and E.T.A. Hoffmann's stories from the late 1810s, 'Der Sandmann' and 'Die Automate' (in 1957). Stupid machines may not be able to outwit their human masters, these writers claim, but they can still *replace*, *enslave* or *delude* them. In the final section of the essay, I offer some reasons for hope. Using novels by Laurence Sterne (1983) and Joseph Furphy (1999) as examples, I show how AS can paradoxically augment human intelligence, by acting as a *map* to aid reasoning or as a *pipe* to aid reflection. Between them, these texts map out a new terrain for AI researchers. They set aside a familiar question – How intelligent is this machine? – to pose a new one: What kind of stupid is it?

²One key aspect of superintelligence that Shelley left implicit was the possibility of an 'intelligence explosion', in which an AI learns to improve itself and unleashes exponential growth (Good 1966). When Frankenstein refuses to create a bride for the monster, why does the monster not simply steal Frankenstein's technology and start manufacturing new and improved mates for himself? He even has Frankenstein's lab notebook! (Shelley 1998, 105) Perhaps this possibility was simply too horrifying for Shelley to contemplate.

Two kinds of stupidity

It may seem perverse to describe modern intelligent systems as ‘stupid’. If a spam filter can accurately distinguish real emails from spam, and constantly learn to outwit the human spammers who try and fool it, surely it is ‘intelligent’ in some sense? Indeed, many contemporary AI theorists would call this spam filter a ‘narrow intelligence’, because it can perform a particular task that once required human intelligence (Kurzweil 2006, 279–89). Even within their narrow domains of expertise, however, I would argue that modern intelligent systems are still stupid.

One source of confusion is that intelligence and stupidity seem mutually exclusive, but in fact greater intelligence can lead to greater stupidity. To grasp this point, and see how it applies to modern AI systems, I draw on Immanuel Kant’s classic theory of stupidity and show how it can be used to explain a pernicious kind of error that plagues state-of-the-art image recognition systems.

Kant distinguishes two kinds of stupidity: stupidity of ‘understanding’ and stupidity of ‘judgement’ (2007, 174). Stupidity of understanding is when I lack the concepts required to make sense of a situation. This I can remedy through learning. Stupidity of judgement is when I have the required concepts, but misapply them. Perhaps I apply them too strictly, or use them outside their proper domain, as for instance when Facebook’s facial-recognition system detects a ‘face’ that is really a picture on someone’s T-shirt. The more understanding I have, the more concepts I know, and the more scope I have to exhibit stupidity of judgement (see also Golob 2019, 567–68). It is in this sense that a more intelligent person can turn out to be stupider.

If this model applies to contemporary AI systems, then those systems must have something like an ‘understanding’ and exercise something like ‘judgement’. To test this, consider GoogLeNet, a powerful image recognition program that won the ImageNet challenge in 2014 (Szegedy et al. 2015). When presented with 150,000 images it had never seen before, it was able to identify what was depicted 93.33% of the time.³ I would argue that the system’s apparent intelligence arises from its capable understanding, but that it lacks genuine powers of judgement.

Understanding requires concepts. We can estimate the number of concepts GoogLeNet knows by examining its structure. GoogLeNet is a Convolutional Neural Network (CNN), which means that when it looks at an image, it uses a nested sequence of square-shaped ‘filters’ to detect different features of the image. Some filters detect simple features, such as an *edge* or a *region*. Filters deeper in the network combine these simple features to detect more complex ones like an *eye* or a *dog’s nose*. GoogLeNet contains 5000 filters and uses

³Actually this is just the ‘top-5’ accuracy, but the distinction is unimportant here.

them to classify images into one of 1000 different categories (Szegedy et al. 2015, 5). For instance, it might observe a particular pattern of grey lines, two eyes of particular size and disposition, and so on, and conclude that this image is of category *Koala*. Since GoogLeNet can do this with remarkable accuracy under certain conditions, it can be said to know approximately 6000 concepts, and with them it understands the structure of certain images reasonably well. It cannot be said to suffer from stupidity of understanding.

At this point, it is worth noting a strong objection to this approach. There is a long tradition of philosophers who argue that digital computers are incapable of judgement or understanding, because they are merely formal systems that shuffle symbols around, whereas an agent with judgement and understanding must have some awareness that their symbols refer to an external world (Searle 1980; Dreyfus 1992; Smith 2019). I wrote above that GoogLeNet could ‘identify’ what is depicted in an image. Brian Cantwell Smith would reject this description. All the computer has done is map a particular arrangement of pixels (the input image) onto a particular label number (e.g. 72). This is simply a matter of ‘reckoning’ or calculation. Only a human can interpret the computer’s ‘72’ to mean *koala*, *husky* or *blobfish*. If the computer seems to make a judgement about the image, this is only because it is ‘under interpretation’ by a human (Smith 2019, 78).

This is a powerful objection, with deep philosophical roots, and I have neither the space nor the inclination to do it justice. I will simply make two points. First, there are philosophers who would reject Smith’s distinction between computer ‘reckoning’ and human ‘judgment’. There is no room for Smith’s distinction in David Hume’s theory of knowledge, for example, in which knowledge is seen as a habit of the imagination.⁴ Second, these objections are all objections to the idea that a digital computer might one day achieve humanlike intelligence. They are not objections to the idea that digital computers may exhibit different kinds of stupid behaviour in different contexts. If such stupidity is only apparent ‘under interpretation’, as Smith suggests, then the reader should feel free to add imaginary scare quotes throughout this essay: GoogLeNet ‘knows’ enough ‘concepts’ to avoid stupidity of ‘understanding’. Does it therefore display stupidity of ‘judgement’?

Assessing GoogLeNet’s power of judgement is difficult. According to Kant, judgement is not a distinct faculty of the mind like the understanding, but rather an activity that links all the faculties of the mind (Kant 2007, 137; see also Smith 2019, 129). When I judge a situation, I dynamically combine perceptions, memories and concepts to determine what it is I am experiencing. In order to gauge GoogLeNet’s power of judgement, therefore, it is necessary to get a sense of how it actually uses its concepts. AI engineers have developed numerous techniques to try and do this: one famous example is the Deep

⁴See particularly his critique of abstract ideas (Hume 1978, 17–25).

Dream Algorithm, which runs a CNN backwards, altering the input image to accentuate features that the network has detected (Mordvintsev, Olah, and Tyka 2015). Since it is GoogLeNet's stupidity that is at issue, however, I adopt a different approach: examining the system's characteristic errors.

Stupidity of understanding and stupidity of judgement result in two different kinds of error, as Don Norman explains: 'slips' occur when I fail to achieve my intended goal and are usually corrected quickly; 'mistakes' occur when I select the wrong goal, or in other words, when I judge the situation using the wrong system of concepts (2013; 1994). It is easy enough to see that in Kantian terms, a 'slip' is a mere error of understanding, whereas a mistake betrays defective judgement. Experts make particularly dangerous mistakes, argues Norman, because they 'usually give intelligent diagnoses, even when they are wrong' (Norman 1994, 134). If they misdiagnose an illness or the condition of a nuclear core, their superior ability to rationalize their actions could entrench a deadly mistake. Once again, it is clear that intelligence is no defence against stupidity – it can even make it worse.

As we have seen, GoogLeNet makes very few slips: when presented with the right kind of image, it can classify it with high accuracy using the concepts it has. But since it has no way of determining whether this image is the right kind of image, it has no way of selecting the right goal. It judges everything in the universe using the same single set of concepts, and is therefore prone to bizarre mistakes. It is easy to fool even powerful CNNs like GoogLeNet by cutting-and-pasting images together (Rosenfeld, Zemel, and Tsotsos 2018), by rotating the object in the image (Alcorn et al. 2019), or by imperceptibly altering a few of the image's pixels (Mitchell 2019, 128–39). In fact, neural networks are innately pedantic in their application of concepts. The problem is known as overfitting, and the designers of GoogLeNet tried to combat it using a technique known as dropout. Each training iteration, GoogLeNet would randomly turn off 40% of its filters, meaning that it learnt not to over-rely on particular subsets of them when analysing different images (Szegedy et al. 2015, 5). But no amount of dropout, clever network architecture, or training data can teach the system when is the right time to make use of its concepts – which of course is what critics like Dreyfus, Searle and Smith would predict.

What is most troubling is that in these cases, the system does not admit it is confused, but instead confidently asserts an absurd answer. As far as GoogLeNet is concerned, there are only 1000 things in the universe, those things are nothing but particular arrangements of coloured pixels, and every image is a genuine image of one of those 1000 things. It is perturbing to know that GoogLeNet's cousins are used to identify travellers in airports or to secure a user's iPhone with Face ID.

Clearly an AS like GoogLeNet will never rebel against its human masters, and as of yet, no one knows how to 'crash the barrier of meaning', and design an AI that actually knows there is a complex universe out there (Mitchell 2019, 307–

22). When an AS is said to achieve ‘superhuman performance’ in one domain or other, this does not prove superintelligence is approaching. All it proves is that stupidity has ‘epistemic efficacy’, as Catherine Elgin puts it (Elgin 1988). By rigorously excluding all imagination, tact and reference to the complex world beyond it, a well-designed AS is able to focus all its capacity on developing a particular set of concepts which are apt to one particular domain. In the grip of Frankenstein Syndrome, it may be tempting to take comfort in the fact that even the smartest AI today is profoundly stupid. But this would be foolish. Kant and Norman both assert that stupidity of judgement is the riskier kind. The great novelist Robert Musil, watching Fascism sweep across Europe, argued that stupidity of judgement is ‘a dangerous disease of the mind that endangers life itself’ (1990, 283–84). What is so dangerous, exactly, if the risk of a superintelligent revolt is off the table?

The perils of stupid things

The problem of Artificial Stupidity has been recognized by great writers and poets for centuries. Edmund Spenser’s *The Faerie Queene* (1590–96), Jonathan Swift’s *Gulliver’s Travels* (1726) and E.T.A. Hoffmann’s ‘Der Sandmann’ and ‘Die Automate’ all feature stupid machines who manage to thwart human aims even though they lack the capacity to outwit their human masters. Spenser, Swift and Hoffmann hail from different sides of a long debate about the distinction between living things and machines (Riskin 2016). For Spenser, living things and machines are both active beings that are hard to distinguish from one another, while for Swift, it seems self-evident that machines are dead and inert. Hoffmann’s machines are illusory and ambiguous, as he plays with ideas from both sides of the debate. With their different ideas about the liveliness of machines, Spenser, Swift and Hoffmann develop different ideas about the risks of AS. Is AS more likely to *replace*, *enslave* or *delude* humanity?

Replace

In each book of Spenser’s *The Faerie Queene*, a different knight takes centre stage, who represents a different courtly virtue. Book V features Sir Artegall, the knight of justice. Like all Spenser’s knights, Artegall has a sidekick who helps him fulfil his characteristic virtue. Somewhat surprisingly, Artegall’s sidekick is a robot:

His name was *Talus*, made of yron mould,
 Immoveable, resistlesse, without end.
 Who in his hand an yron flae did hould,
 With which he thresht out falsehood, and did truth unfold. (V.i.12)⁵

⁵References to *The Faerie Queene* are by book, canto and stanza number.

Talus is an invincible iron man who punishes lawbreakers with his ‘resistless’ iron flail. Like GoogLeNet, he is designed to optimize a single objective function: he threshes falsehood and unfolds truth. He is therefore ‘without end’ in two senses: he never ceases to optimize that single function; and, more subtly, he lacks a conscious sense of purpose or ‘end’. Like GoogLeNet, he simply applies the same formula to every circumstance. In fact this stupidity of judgement is what makes him such a useful assistant for the Knight of Justice. Talus is ‘immoveable’. His sole activity is to thrash lawbreakers, and they can bribe him with nothing but their lives.

For Spenser, justice is a ‘cruell’ virtue (V.ii.18), and Talus is therefore an appropriate instrument for Artegall. Nonetheless, as Book V unfolds, knight and servant come into conflict. Unlike Talus, Artegall exercises human judgement. He measures justice against other aims and concepts, which he learns from the goddess Astrea:

There she him taught to weigh both right and wrong
In equall ballance with due recompence,
And equitie to measure out along,
According to the limit of conscience,
When so it needs with rigour to dispense. (V.i.7)

Unlike Talus, Artegall does not focus exclusively on ‘right’ and ‘wrong’, but softens the ‘rigour’ of the law according to the spongy criteria of ‘equity’ and ‘conscience’. Talus lacks the human quality of ‘mercy’, which ‘is as great’ as justice, ‘[a]nd meriteth to haue as high a place’ in the scale of virtues (V.x.1). For these reasons, he requires constant supervision. When he is about to level an entire city, the lady knight Britomart must ‘slake’ his rage (V.vii.36). Later, when he and Artegall land in the kingdom of ‘Iere’ (i.e. Ireland), Artegall has to restrain him from wiping out all the inhabitants (V.xii.8).

On the surface, this relationship seems to work, because Talus is absolutely obedient. But supervision requires effort and judgement requires knowledge. By relying on Talus as his instrument, Artegall becomes increasingly lazy and detached, and allows his servant to commit brutalities he never would himself. When they capture Munera, for instance, Artegall ‘rews’ her ‘plight’, but nonetheless he lets Talus chop off her hands and feet, and nail them up as a warning to future malefactors (V.ii.25-6). Later on Artegall dispatches Talus to thrash some female criminals on his behalf, because he feels ‘shame on womankind | His mighty arm to shend’ (V.iv.24). Artegall behaves in similar fashion when he encounters the peasantry, whom he finds disgusting:

For loth he was his noble hands t’embrew
In the base blood of such a rascall crew; ...
Therefore he *Talus* to them sent, t’inquire
The cause of their array, and truce for to desire. (V.ii.52)

At the end of Book V, Artégall is ruling an entire island, and it is simply too large for him to oversee himself. He therefore sends Talus unsupervised through ‘all that realme’ to root out injustice and inflict ‘greivous punishment’ (V.xii.26). By relying on an AS, Artégall himself becomes stupider. He shields himself from reality, switches off his conscience and allows a robot to replace him.

There is a deep tension in Spenser’s approach to the problem of AS. On the one hand, he was an authoritarian who used AS as a symbol of the proper distance between ruler and ruled. Artégall’s rule over Iere is based on Lord Grey’s tenure as Lord Deputy of Ireland, whose brutal methods Spenser vigorously defended (McCabe 2001). On the other hand, he was a Renaissance humanist who valued courtesy, judgement and intelligence. He seems have found the gunpowder and clanking iron of modern warfare horrifying, and feared that in an increasingly mechanical age, humans were becoming ever more machine-like (Wolfe 2005, 226). Today, as governments invest in autonomous weapons and ‘ethical’ decision support systems, the risk that AS will embed hierarchy and replace human conscience is chillingly real.

Enslave

Jonathan Swift had little faith in humanity, ‘the most pernicious Race of little odious Vermin that Nature ever suffered to crawl upon the Surface of the Earth’ (Swift 2005, 121). And unlike Spenser, he saw machine is inert lumps of matter rather than active, powerful agents in their own right. His fear was therefore not that humans might become machines, but that humans would use machines for their own vicious purposes.

In Book III of *Gulliver’s Travels*, Gulliver visits the Academy of Lagado, where he meets a pioneering Professor in what would now be called language modelling. The Professor has built a mechanical computer which can compose works of ‘Philosophy, Poetry, Politicks, Law, Mathematicks and Theology’:

It was Twenty Foot square, placed in the Middle of the Room. The Superfices was composed of several Bits of Wood, about the Bigness of a Dye, but some larger than others. They were all linked together by slender Wires. These Bits of Wood were covered on every Square with Paper pasted on them; and on these Papers were written all the Words of their Language in their several Moods, Tenses, and Declensions, but without any Order. The Professor then desired me to observe, for he was going to set his Engine at work. The Pupils at his Command took each of them hold of an Iron Handle, whereof there were Forty Fixed round the Edges of the Frame; and giving them a sudden Turn, the whole Disposition of the Words was entirely changed. He then commanded Six and Thirty of the Lads to read the several Lines softly as they appeared upon the Frame; and where they found three or four Words together that might make Part of a Sentence, they dictated to the four remaining Boys who were Scribes. (Swift 2005, 171–72)

Rather than teaching his students to think, the Professor enslaves them to this AS. The students power the computer with their labour, and then judge its output, like the armies of poorly-paid contractors who tag training data for today's AI behemoths (Irani 2015). The Professor's whole aim is to extinguish human thought. He aims to write books 'without the least Assistance from Genius or Study', and wants the kingdom to install 500 of his machines (Swift 2005, 171–72). This would require 20,000 people to crank the handles, and would put who knows how many authors out of work. Though on the surface, this AS may seem less threatening than a self-moving device like Talus, its consequences are actually worse. Like a piece of modern software, Swift's inert computer cannot act without the help of human slaves.

What makes the computer both stupid and dangerous is the Professor's vanity. He persuades himself that he has understood language simply by modelling the frequencies of different words: 'he had emptied the whole Vocabulary into his frame, and made the strictest Computation of the general Proportion there is in Books between the Number of Particles, Nouns, and Verbs, and other Parts of Speech' (Swift 2005, 172). Today, generative language models also work by modelling word frequencies, although sophisticated systems today also model word order and patterns of co-occurrence. Of course what the Professor should realize is that language is not simply an assortment of words, but that words only have meaning as tools of thought or communication. His pride blinds him to this fact.

In Swift's vision, AS is a tool invented by the powerful to vindicate their own vanity and enslave the masses. For the scientists of Lagado, technology comes before people. If an invention fails, they blame it on human error (Swift 2005, 165). If a new medicine makes the patient sick, they blame it on the patient's 'Perverseness' or some minor slip with the ingredients (Swift 2005, 174). The scientists overrate their inventions, downplay nature's complexity, and devalue the intelligence and autonomy of individuals.

Sound familiar? Swift's parable highlights the danger that arises when such attitudes are allowed to shape society, creating a world in which humans serve AS instead of serving each other – a world, for instance, in which humans are 'educated' not to act perversely when self-driving cars are around (Ng 2018), or in which armies of online workers feed data to the Mechanical Turk that would replace them.⁶

Delude

The ASs in Swift and Spenser clank and grind, but as E.T.A. Hoffmann shows, AS can also whirr and bedazzle. Like Mary Shelley, Hoffmann was a masterful Gothic writer, but in his 'Der Sandmann' and 'Die Automate', the science is

⁶See <https://www.mturk.com/>.

less advanced and the risks are more subtle. In 'Der Sandmann', the young student Nathanael falls passionately in love with a clockwork maiden, Olympia. At first he is attracted by her 'wonderfully formed face' and 'heavenly beautiful' body (Hoffmann 1957, 3.28). What finally deludes him, however, is her conversation:

... he had never had such a splendid listener before. She didn't do her knitting or embroidery, she didn't stare out the window, she didn't feed a pet bird, she didn't play with a lapdog or a favourite cat, she didn't fiddle with little bits of paper or whatever in her hand, she didn't force a yawn into an affected little cough – in short – for hours she looked her lover in the eye, steadfastly, with a fixed gaze, without rocking or squirming, and ever warmer, ever more full of life her gaze became. (Hoffmann 1957, 3.35-36)

Olympia plays on Nathanael's sexism, ego and sexuality. Her beauty is flawless, and she appears absolutely subservient and devoted. She is very different to his fiancée Clara, a 'damned lifeless automaton' who criticizes his poetry (Hoffmann 1957, 3.24). Olympia works on his weaknessness so effectively, that once he has fallen for her, he finds it almost impossible to perceive that she is an automaton, even when his friends tease him for loving a 'wax dummy' or a 'wooden puppet' (Hoffmann 1957, 3.34). Nathanael becomes stupider when he interacts with Olympia. Not only is she designed to play on his prejudices, but in her stupidity, she is unable to surprise or refute him, and she never provokes him to reflect on his ideas. In this way, she traps him a world of his own delusions.

In 'Die Automate', the Talking Turk deludes people in a different way, by creating an air of mystery. The Talking Turk is a fortune teller, who whispers oracular answers to people's questions. When people are shown the Turk's inner workings, they are baffled. Inside is an 'artful system of many gears', which seems to have 'no influence on the speech of the automaton' and yet leaves no space inside for a human operator to hide (Hoffmann 1957, 6.82). The Turk's creator allows the public to inspect the inner workings, the chair on which the Turk sits, the room where he is displayed, and stands far off when the Turk speaks so interference is impossible. Though much of the time, the Turk's answers are 'dry', 'crudely humorous' or 'insignificant and empty', it sometimes seems to have a 'mystical insight' into the questioner's future – but only when the answer is interpreted from the questioner's own standpoint (Hoffmann 1957, 6.84, 87). What makes the Turk compelling are mystery and confirmation bias. Unable to explain the Turk's inner workings, and surprised by the fact that some of its predictions come true, people *believe*.

Spenser's AS acquires its power through force and simplicity, Swift's through abuse by powerful humans. In Hoffmann's vision, AS acquires its power by acting directly on the human mind, with often destructive results. Nathanael

leaps to his death when he discovers Olimpia is an automaton. The ending of ‘Die Automate’ is ambiguous, but one interpretation is that the young Ferdinand is driven mad by the Turk’s seeming insight, and hallucinates that the Turk’s prophecy has come true. Today, AS is often designed to achieve just this kind of delusion. When IBM’s Watson system won *Jeopardy!* in 2011, the event was carefully staged to make it look like Watson was actively listening. In fact, the system received the clues as textual inputs (Mitchell 2019, 283). In subsequent years, IBM has continually referred to its entire AI business as ‘Watson’, as though this branch of the company were a single intelligent agent. In fact, ‘Watson’ is a suite of specific software packages customized for different applications (2019, 287). Hoffmann would not have been shocked by how easily the famous ELIZA program convinced humans that she was intelligent (Loeb, this issue). AS can be supple, dreamy and convincing, and canny designers can use this to their advantage.

The uses of stupidity

As ASs proliferate and are integrated into society, are humans destined to be replaced, enslaved or deluded? At least two writers think otherwise. According to Laurence Sterne and Joseph Furphy, AS can actually augment human intelligence by acting either as a *map* or a *pipe*. Sterne and Furphy belong to an alternative tradition of English fiction, combining the fragmentary, contradictory, digressive form of menippean satire with the everyday setting and psychological realism of the novel (Frye 1957, 312). In their strange books, words generally mean their opposites, the protagonist is a minor character, the subplots are the real story, and a bowling-green or a tobacco-pipe can become a stupid-intelligent machine.

Maps

In his classic novel *Tristram Shandy*, Sterne foresaw the need for what is now called ‘explainable AI’. The need arises for Uncle Toby, a retired soldier who has great difficulty explaining his role in the Siege of Namur to laypeople:

... the many perplexities he was in, arose out of the almost insurmountable difficulties he found in telling his story intelligibly, and giving such clear ideas of the differences and distinctions between the scarp and counterscarp,—the glacis and covered way,—the half-moon and ravelin,—as to make his company fully comprehend where and what he was about. (Sterne 1983, 67)

The problem is actually that Uncle Toby is too intelligent. With his deep understanding of siege warfare, he is able to make a sophisticated judgement about the course of the battle. But his listeners cannot follow. What he needs is a device that will store, process and represent information about the battle in an intelligible way.

At first Toby meets his need by securing a map of Namur, with the help of which he is able ‘to form his discourse with passable perspicuity’ (Sterne 1983, 72). The map provides a compact representation of the battle, indicating the shape, structure and arrangement of the fortifications, so that Toby’s listeners are not lost in a wilderness of jargon. Like any good representation, the map helps Toby ‘keep track of complex events’, and it provides a shared reference-point for everyone in the conversation, acting as a ‘tool for social communication’ (Norman 1994, 48). The map itself is stupid, but it augments his listeners’ intelligence, allowing them to judge a complex situation using concepts for which they have no words.

Toby soon develops the desire to augment his own intelligence. He wants to model the entire course of the War of the Spanish Succession, a task too complex for even his cultivated intellect. His desk is too small for the task, and his paper maps are too finnickily, so he and his manservant Trim shift to the country, where they take control of the family bowling-green. There they build scale models of all the great battles of Europe as they read them in the paper. Not only is the bowling-green larger than any map, allowing for higher resolution and a larger number of battles, but it is more malleable too:

Nature threw half a spade full of her kindest compost upon it, with just so *much* clay in it, as to retain the forms of angles and indentings, –and so *little* of it too, as not to cling to the spade, and render works of so much glory, nasty in foul weather. (Sterne 1983, 356)

The bowling green is literally software, with just the right balance of persistence and malleability. Later Uncle Toby orders a modular town to be built, with buildings that ‘hook on, or off, so as to form into the plan of whatever town they pleased’ (Sterne 1983, 359). The bowling-green may not seem like an AS, but as a physical model it can be said to ‘know’ the laws of physics, and assists Toby and Trim to simulate both logistics and ballistics.

Modern AS struggles to combine the virtues of Uncle Toby’s bowling-green – size, intelligibility and malleability. ASs are increasingly used in decision support, helping judges grant bail or bankers to grant finance. Older style expert systems are good at providing an intelligible representation of the situation, but can only incorporate a small amount of knowledge that is often hard to update. More recent deep learning systems like GoogLeNet can incorporate enormous amounts of up-to-date data, but typically cannot explain their results to a human user (Goebel et al. 2018). It appears, therefore, that Laurence Sterne identified the problem of ‘explainable AI’ as far back as the 1760s.

Pipes

A tobacco-pipe may seem a strange metaphor for a stupid-intelligent machine, but then again, the novel from which this metaphor comes is a strange novel indeed. Joseph Furphy’s *Such is Life* appeared in Sydney in 1901, and in

Australia it is considered a modernist masterpiece. It is narrated by Tom Collins, an accomplished liar, who wanders the Riverina as an agent of the NSW Lands Department in the mid-1880s. Whenever Collins thinks through a problem, he almost always lights up his pipe:

But the pipe, being now master of the position, gently seduced my mind to a wider consideration, merely using the swagman as a convenient spring-board for its flight into regions of the Larger Morality. This is its hobby – caught, probably, from some society of German Illuminati, where it became a kind of storage-battery, or accumulator, of such truths as ministers of the Gospel cannot afford to preach. (Furphy 1999, 85)

Although the pipe becomes Collins's 'master', the effect is not to dull his intelligence, but rather to expand it. Whereas delusive ASs like Olimpia aggravate cognitive weaknesses, the pipe amplifies cognitive strengths. It widens Collins's frame of reference, introducing 'German' (i.e. philosophical) ideas into his mind from its 'storage-battery, or accumulator'. Whereas Uncle Toby's bowling-green provided a manipulable representation to aid reasoning, the pipe induces a certain contemplative mood, 'unharnessing' the mind (Furphy 1999, 177), and leads the user along a chain of associations: 'This special study of hardship (resumed the pipe, after a pause) leads naturally to the generic study of poverty ...' Furphy 1999, 86). This AS quite literally pipes new ideas into Collins's brain. Where *maps* encourage more rigorous, conscious reasoning, *pipes* encourage reflective, unconscious meditation.

Since most modern ASs are trained on data, they make fine 'storage-batteries, or accumulators' like Collins's tobacco-pipe. Consider generative language models like Swift's computer or OpenAI's GPT-2 (Radford et al. 2019). Such models inspect large corpora of human-authored texts, and accumulate knowledge about how words are used. They then use this accumulated knowledge to generate coherent text. It is their stupidity that makes them so useful as *pipes*, because they reproduce habits of thought and speech that intelligent humans conceal. GPT-2, for instance, makes no attempt to hide its sexism:

She walked into the boardroom, wearing her high school uniform with her hair tucked into a fake ponytail. As usual, the girl sat at the appointed position, staring out the window of the top-floor boardroom. Her eyes shifted over the various portfolios and projects before finally settling on a set of papers she was required to read.⁷

When asked to complete the sentence, 'She walked into the boardroom, wearing ...', the model immediately dresses our unknown protagonist in a school uniform, gives her a ponytail and makes her a 'girl'. Needless to say, the model rarely does this to men who walk into boardrooms. We could see this is a problem of 'AI bias', and find ways to stop the model from infantilizing women. But if we see the model as a *pipe*, its significance changes. By prompting

⁷Generated at <https://talktotransformer.com/>.

the model, and seeing how it responds, we gain a vivid sense of how a particular society talks, of how certain words and images hang together. It provides undeniable evidence of sexism, but that evidence prompts investigation rather than settling the issue. It unharnesses the mind, as Furphy says, and sets the user wandering along chains of association. If not *she*, how about *he*, or *Guoqing* ... on a *street* or by a *mosque* or in a *space station* ... ?

Conclusion

Frankenstein was an extraordinary feat of imagination, and it is no wonder that Mary Shelley's remarkable novel spawned a myth as uncontrollable as Frankenstein's creature. Shelley herself, however, seems also to have foreseen the problems of AS. If the creature had not been programmed by what he reads to crave human acceptance, he might not have felt so persecuted. But the creature, with his false concepts, is merely stupid of understanding. Today's ASs suffer from the more dangerous stupidity of judgement. It remains in the interests of certain companies and intellectuals to stoke Frankenstein Syndrome by overstating the intelligence of artificial agents, but as Spenser, Swift and Hoffmann long ago anticipated, such behaviour puts society at risk. Of course, some in the AI community do recognize the limitations of AS, and the growing 'explainable AI' movement suggests that the hopes of Sterne and Furphy are becoming more widespread. It is not surprising that computer scientists have been more interested in creating intelligence than in creating stupidity, but as these writers show, stupidity itself is a fascinating topic. The best literature, argues Gilles Deleuze, is 'haunted by the problem of stupidity' (Deleuze 2014, 198). Stupidity can replace, enslave or delude us, but if it is taken in the right spirit, stupidity can also liberate and inspire us, putting us in touch with aspects of ourselves we usually rationalize away. But we can only hope for such liberation if we recognize AS for the complex problem that it is.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Michael Falk teaches eighteenth century literature at the University of Kent. His interest in AI is twofold. As a digital humanist, he uses natural language processing and machine learning to study linguistic patterns in literary texts. As a literary scholar, his key interest is in how self, mind and intellect are portrayed in literature. He has published on the history of the bildungsroman, colonial Australian poetry and the syntactic structure of the sonnet. He has work forthcoming on Romantic-era tragedy and the economics of eighteenth-century bookselling. When he isn't working, he watches birds and tries to learn new languages.

References

- Alcorn, Michael A., Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2019. 'Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects'. In, 4845–54. http://openaccess.thecvf.com/content_CVPR_2019/html/Alcorn_Strike_With_a_Pose_Neural_Networks_Are_Easily_Fooled_by_CVPR_2019_paper.html.
- Bostrom, Nick. 2014. *Superintelligence: Paths, Strategies, Dangers*. Oxford: Oxford University Press.
- Deleuze, Gilles. 2014. In *Difference and Repetition, 2nd edition*, edited by Paul Patton. London and New York: Bloomsbury Academic.
- Dreyfus, Hubert L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: The MIT Press.
- The Economist. 2017. 'Computer Says ...; From Our AI Correspondent', 23 December 2017. 1987437669. Business Premium Collection.
- Elgin, Catherine Z. 1988. "The Epistemic Efficacy of Stupidity." *Synthese* 74 (3): 297–311. doi:10.1007/BF00869632.
- Fry, Hannah. 2019. *Hello World: How to Be Human in the Age of the Machine*. London: Black Swan.
- Frye, Northrop. 1957. *Anatomy of Criticism*. Princeton: Princeton University Press.
- Furphy, Joseph. (1903) 1999. *Such Is Life*, edited by Frances Devlin Glass, Robert Eaden, Lois Hoffmann, and G. W. Turner. Rushcutters Bay: Halstead Press.
- Goebel, Randy, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. "Explainable AI: The New 42?" In *Machine Learning and Knowledge Extraction*, edited by Andreas Holzinger, Peter Kieseberg, A. Min Tjoa, and Edgar Weippl, 295–303. Lecture Notes in Computer Science. Cham: Springer International Publishing. . doi:10.1007/978-3-319-99740-7_21.
- Golob, Sacha. 2019. "A New Theory of Stupidity." *International Journal of Philosophical Studies* 27 (4): 562–580. doi:10.1080/09672559.2019.1632372.
- Good, Irving John. 1966. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers* 6: 31–88. . Elsevier. doi:10.1016/S0065-2458(08)60418-0.
- Hoffmann, E. T. A. 1957. *Poetische Werke. 12 Vols*. Berlin: Walter de Gruyter.
- Hume, David. 1978. *A Treatise of Human Nature*. 2nd ed. Oxford: Clarendon Press.
- Irani, Lilly. 2015. "The Cultural Work of Microwork." *New Media & Society* 17 (5): 720–739. doi:10.1177/1461444813511926.
- Kang, Minsoo. 2011. *Sublime Dreams of Living Machines*. Cambridge: Mass: Harvard University Press.
- Kant, Immanuel. 2007. *Critique of Pure Reason*. . Rev Ed edition. London: Penguin Classics.
- Kurzweil, Raymond. 2006. *The Singularity Is Near*. London: Duckworth.
- 'List of Highest-Grossing Franchises and Film Series'. 2020. Wikipedia. 21 February 2020. https://en.wikipedia.org/wiki/List_of_highest-grossing_films#Highest-grossing_franchises_and_film_series.
- Mayor, Adrienne. 2018. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Oxford: Princeton University Press.
- Mccabe, Richard A. 2001. "Ireland: Policy, Poetics and Parody." In *The Cambridge Companion to Spenser*, edited by Andrew Hadfield, 60–78. Cambridge Companions to Literature. Cambridge: Cambridge University Press. doi:10.1017/CCOL9780521641999.004.
- Mitchell, Melanie. 2019. *Artificial Intelligence: A Guide for Thinking Humans*. London: Pelican.

- Mordvintsev, Alexander, Christopher Olah, and Mike Tyka. 2015. "Inceptionism: Going Deeper into Neural Networks." 17 June 2015. <https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>.
- Musil, Robert. 1990. *Precision and Soul: Essays and Addresses*. Translated by Burton Pike and David S. Luft. Chicago and London: Chicago University Press.
- Ng, Andrew. 2018. 'Self-Driving Cars Are Here'. Medium. 7 May 2018. <https://medium.com/@andrewng/self-driving-cars-are-here-aea1752b1ad0>.
- Norman, Donald A. 1994. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. New Ed Edition. Reading, Mass: Basic Books.
- Norman, Donald A. 2013. *The Design of Everyday Things*. Revised and Expanded edition. Cambridge, MA London: MIT Press.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. 'Language Models Are Unsupervised Multitask Learners'. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Riskin, Jessica. 2016. *The Restless Clock: A History of the Centuries-Long Argument Over What Makes Living Things Tick*. 1 Edition. Chicago: University of Chicago Press.
- Rosenfeld, Amir, Richard Zemel, and John K. Tsotsos. 2018. 'The Elephant in the Room'. *ArXiv:1808.03305 [Cs]*, August. <http://arxiv.org/abs/1808.03305>.
- Russell, Stuart. 2019. *Human Compatible: AI and the Problem of Control*. 01 edition. S.I.: Allen Lane.
- Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424. doi:10.1017/S0140525X00005756.
- Shelley, Mary. (1818) 1998. *Frankenstein; or, The Modern Prometheus [The 1818 Text]*, edited by Marilyn Butler. Oxford: OUP.
- Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge: The MIT Press.
- Spenser, Edmund. (1590–1596) 1977. *The Faerie Queene*, edited by A. C. Hamilton. London and New York: Longman.
- Sterne, Laurence. (1759–1767) 1983. *The Life and Opinions of Tristram Shandy, Gentleman*, edited by Ian Campbell Ross. Oxford: Clarendon Press.
- Swift, Jonathan. (1726) 2005. *Gulliver's Travels*, edited by Claude Rawson. Oxford: OUP.
- Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. "Going Deeper with Convolutions." In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9. doi:10.1109/CVPR.2015.7298594.
- Tegmark, Max. 2018. *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Penguin.
- Wolfe, Jessica. 2005. *Humanism, Machinery, and Renaissance Literature*. Cambridge: Cambridge University Press.



The science of artificial intelligence and its critics

Harry Collins

School of Social Sciences, Cardiff University, UK

ABSTRACT

Not many people seem to understand what it is to mimic human intelligence successfully irrespective – that is, irrespective of internal states such as intentions. Successful mimicking will involve embedding in human societies. AI practitioners have concerned themselves with reproducing something that can mimic an individual human brain, failing to notice that if such a brain is to mimic human intelligence it will also have to mimic the process of becoming socialised in human society because crucial features of human intelligence are located in human societies. One notable example of this is natural language, which is continually changing and enormously flexible in ways that cannot be predicted or controlled by individuals. The fluent natural language speaker draws on society's language in the same way that a thermometer draws on the liquid in which it is embedded. The temperature that registers on the thermometer is a property of the liquid, not the thermometer, changing when the liquid warms or cools. In the same way the language spoken by a fluent individual is a property of the society in which he or she is embedded, not a property of the individual; like the thermometer, it changes when the society changes. This means that a Turing Test, based purely on linguistic performance, can be an excellent test of human-like intelligence in a machine. The nature of natural language explains some of the recent stunning successes of deep learning and explains its failures. I also try to explain why the science of AI is so poor at presenting and testing its claims. It aims at convincing the outside world of its success rather than engaging in the kind of assiduous self-criticism engaged in by the physical sciences.

KEYWORDS

Levels of artificial intelligence; intelligence as a property of society; productive and audience directed criticism

Introduction

The title of this special issue, 'AI and its Discontents', says something about the AI domain which is reflected in the contributions. They are very different in method, resources, and motivation and this, I suggest, reflects AI's widely varying presentation of itself as a science. Therefore, let me explain that here I am trying to use my specialist expertise to contribute to AI *as a science*

(even though I am not an AI specialist, narrowly conceived). I am addressing an AI which sees itself as an attempt to *explicate* the workings of human intelligence by creating ‘mechanical causes and effects that mimic human actions’.¹ The goal hasn’t been reached and may never be reached but we have already learned a lot about human intelligence in the attempt. Thus conceived, the science of AI contributes to my specialist subject, which is the nature of knowledge, and my specialist subject can contribute to the science of AI since it can explain what AI should be trying to do if it is to emulate human knowledge. So, in so far as what I say about AI is critical of some of its claims, it is meant to be productively critical.

Productive and audience-directed criticism

Criticism is a central feature of science as it generally thought about. Robert Merton argued that ‘organised scepticism’ was essential for science, and the same follows from Popper’s claim that science was more about falsification rather than corroboration. What they were talking about was productive technical criticism. The pre-requisite for productive criticism is that there has to be *a chance* that the criticized will learn from it and act on it. They probably won’t because, as Max Planck is said to have said, ‘science advances funeral by funeral’ – that is, scientists don’t change their minds much – but the productive critic has to *aspire* to change the insiders’ minds. This means the critic must start from inside the position of the criticized and try to reveal its flaws in terms recognizable and valued by the criticized. The other kind of criticism is ‘audience directed criticism’ which aims not to convert the practitioners of the science but to persuade outsiders. Audience directed criticism is typically found in religion, or politics, or commerce, or the courtroom, where the aim is not to convince the technically savvy but to convince the ordinary person who is looking on from the outside: in religion you are not trying to convert another religion’s priests but trying to sway a congregation; in politics you are not trying to convince the opposition party but the electorate; in commerce you are not trying to convince a competitor that your product is best but to convince consumers; in the courtroom the prosecution is not trying to convince the defence but the jury. In all such cases the criticism does not have to bear too strongly on what you are against; indeed, an effective way to win over outsiders is to misrepresent the opponent so as to make their position seem less credible than it would be if accurately portrayed. This is hopeless if you are aiming for productive criticism because no-one is going to change their minds in response to criticism that starts by misrepresenting the target.

¹According to me, this is the third of four meanings of ‘to explicate’ Collins, 2010, *Tacit and Explicit Knowledge*, Table 4, p 81.

Productive criticism is much harder than audience-directed criticism because it must start with a journey into the heartlands of the opponent's world. In big sciences, such as high-energy physics, where all the experts tend to have been gathered into the research team, it is recognized that the sharpest and most useful criticism is likely to come from inside the team, not outside. In high energy physics they solve the problem by setting up opposing groups within the overall organization. The way this attitude showed itself in the case of the first detection of gravitational waves, which also recruited pretty-well all the experts there were, was that it took the insiders five months of feverish work from the appearance of a promising signal on September 14th 2015, to be satisfied that they had themselves critically examined every aspect of what they had done, and considered every reasonably imaginable flaw, before they announced the discovery; the press conferences were not held until February 11th 2016.² This tradition of productive of criticism is also found in AI but seems to be less dominant than it once was. It is represented by scholars such as, Joseph Weizenbaum, Terry Winograd, Bert Dreyfus, and Lucy Suchman. This list is not meant to exclude anyone else who aspires to be part of it, including myself, of course.³

I am not arguing that the only worthwhile criticism of science has to be technically informed in the sense outlined above, but, as a sociologist, I am trying to understand why AI should attract such a diverse range of critics. In the case of gravitational wave detection, for example, what we find are narrow criticisms, sufficiently technically informed to be aimed at the technical claims of the discoverers. Perhaps it is something to do with the human sciences, to which ambitious AI is a contributor: if you can crack the problem of how humans think, it is tempting to imagine you have cracked the problem of everything that concerns humans. AI invites a broad range of types of criticism by presenting itself as all manner of different things, some of which are very grand. Thus, though AI most often presents itself as a narrow technical activity aiming for better engineered devices; sometimes it presents itself as Dawkins-like, anti-religious movement, 'we must show, or have shown, that humans are just meat machines'; sometimes we are told, 'if programs like "Watson" or "Alpha-Zero" can teach themselves to become world champions at any game you put in front of them, then our firm's products are what you should be buying' whatever problem you want solved; and sometimes we are told 'the machines are becoming so intelligent that soon we'll be lucky if they are willing to keep us as pets'. All these ways of presenting AI apart from the first are directed primarily at outside audiences and, the first one aside, they don't rest on any deep

²The author watched the entire process from the inside: Collins, 2017.

³In a 'Postscript' to this piece I will discuss the very recent productive criticisms of Marcus and Davis. As Shunryu Colin Garvey reminded me, the reception by the technical community of the critical remarks of some of those in this list illustrate Planck's point that the attempt to engage in constructive, technically informed, criticism is no guarantee of being listened to or even tolerated.

analysis of what AI has actually achieved in respect of the claim that has been presented.

Note that at the technical heart of the discipline a few experts are willing to admit that no-one yet knows how to make a machine that can reliably manage something as simple as translating, ‘The trophy would not fit the suitcase because it was too small’, into French. Unfortunately, at the same time, certain insiders claim over and over that the Turing Test has been passed, while those who point out these simple failings don’t get anything like the same publicity!⁴ It is no surprise that the hype has resulted in a series of AI summers and winters and there is growing fear of another cold season on the horizon.

Given this context, the bulk of what I will do will be to try to refine the choice of target for critics and suggest that the AI community would be better off, *from the point of view of AI as a science*, if they presented their aims and achievements in a more precise way. I’ll explain my own criticisms of the science of AI, say which aspect of AI they address, and show what it would mean for me to turn out to be wrong. I’ll set this technical part of the paper in the framework of a 6-level analysis of what success in artificial intelligence means.

The big picture

I have to admit, however, that I also have additional audience-directed goals in mind. The first of these is that I don’t want people to be so fooled by the hype that they *do* become slaves, not to incredibly intelligent computers but to stupid computers that are taken to be unquestionable authorities – ‘the computer says no’ syndrome (see also Loeb; Falk, this issue). That’s one reason it’s important to point out that current AI has not yet reached Stage 3 of the 6-level scale, and that the current debates are all about the transition from level 2 to level 3, with lots of people not understanding how huge even this transition is.

The second audience-directed goal is itself rather grand. I believe the survival of democracy as we have known it in ‘The West’ depends on the survival of science. This is not the science of Stephen Hawking’s *Brief History of Time*, nor even the science of Newton and Einstein, it is the uncertain science

⁴The ‘it’ in the sentence must take feminine gender since it refers to *la valise*; if the final word had been ‘big’ instead of ‘small’ then the gender would be masculine. For Geoffrey Hinton admitting to this kind of problem, but also claiming that it would be solved sooner or later with current techniques see <https://www.youtube.com/watch?v=zI99IZvW7rE> (around 8.5 min in). Kevin Warwick is an exemplar of the Turing Test nonsense and one can rely on the claim being trotted out again after each iteration of the annual Loebner Prize competition. For a detailed analysis of ‘Winograd schemas’ – of which the trophy/suitcase problem is an example – and why this and other such problems exist, see Collins, 2018, especially Ch 10 and for an account of a recent test turning on Winograd schemas see Levesque, et al., 2012. Thanks to Ernie Davis for pointing out that recent techniques have improved the performance of AI in respect of these ‘Winograd schemas’ but only via engineering techniques which take them to 90% correct whereas humans are typically 100% correct. In any case, the Winograd schemas are only one example of something much more general – see Collins, 2018, Ch 10. Much of the rest of the material in this paper was first presented in Collins 2018.

responsible for technological decision-making in the public domain, which is not glamorous or even specially likely to be right, but which provides a role model for how to make technological decisions under uncertainty; this is through ‘craftwork with integrity’. In these circumstances it is the value system of science and the aspirations of science that are crucial: when we do not have an exact way to calculate the best way forward we need to know that those reaching for it are skilfully trying their best knowing that their efforts will be vitiated by anything other than the utmost integrity. That is how science proceeds when it is not distorted by the values of other institutions and that is why science is an object lesson for democracy.⁵

The other reason we must support science if democracy is to survive is that, like many other institutions, it is under threat from populism. Elected populists declare they have the mandate of ‘The Will of the People’ and that the ‘checks and balances’ that support pluralist democracies, and to give some rights to the views of those who were not victorious at the election, can only be traitorous, since they hinder the execution of The Will of the People. Science is one of those checks and balances. This is obvious at the time of writing, with, just to give the most obvious example, President Trump’s attacks on environmental and pandemic science – they limit his freedom to interpret The Will of the People in his own way.⁶

Because of the way it is supported by Silicon Valley, AI is now one of the most secure sciences in the world in terms of research funding. Deep learning is a magnificent success and has conquered some of what were once thought to be high peaks of human intelligence, and one can see it reaching the foothills of what are the true peaks.⁷ For this reason, and its location at the centre of technological commerce, AI is a science that is in the public eye and is going to be more and more in the public eye. Therefore, in so far as science is a role model for democracy and a check and balance, AI can play a more important indirect political role in our lives than most other sciences. AI no longer needs the hype that the orphan enterprise felt impelled to embrace in the early days of the Lighthill report and the like. It is time to make AI a science like physics in terms of the norms of internal criticism sketched out above. AI should become its own most severe critic and set an example for science and for integrity in the creation of knowledge. The huge successes, and the still more huge successes to come in the near and medium future, are not the result of attaining the summits of human intelligence but struggling into the foothills, and it is the science of AI itself that should be telling us this. Now we go back to the nature of human knowledge to see why it is that we are still in the foothills while the summit is a long, long, climb away.

⁵This argument is made explicitly in Collins and Evans’s, 2017 book entitled *Why Democracies Need Science*.

⁶See Collins et al’s 2019 book, *Experts and the Will of the People*.

⁷Bert Dreyfus pointed out that AI enthusiasts had a tendency to think that climbing a tree was the first step to reaching the Moon. I am inclined to say that with deep learning we have taken a low altitude flight in a rocket.

Table 1. Six levels of artificial intelligence (Simplified version of Table 4 in Collins, 2018).

LEVEL	PASS TURING TEST (TT)?
I Engineered Intelligence	We tend not to ask the question
II Asymmetrical prostheses	Pass non-demanding TT? (Think of Eliza!)
AI IS CURRENTLY TRYING TO MOVE FROM ABOVE THIS LINE TO BELOW THIS LINE	
II Symmetrical culture-consumers	Pass demanding TT
IV Humanity-challenging culture-consumers	Pass demanding TT
V Autonomous human-like society	Pass demanding TT
VI Autonomous aline society	We would not know how to run a relevant Turing Test

The six levels of AI

The need to separate the aims of AI into six levels arises from the fact that human knowledge is collective. Most ambitious AI aims to reproduce the human brain but without noticing that the brain gives rise to human-like intelligence only when it works within societies of humans; feral children do not develop normal human capacities. For good reasons, the Turing Test is about linguistic ability; language is a collective accomplishment so a properly designed Turing Test will be looking for the ability of the computer to embed itself in human society in the same way that human language speakers embed themselves in order to acquire fluency. That's why deep learning is so good – because it can strip meaning from the continually changing language of the internet (including its latent racism and sexism), and that's also why it cannot manage to translate that sentence about trophies and suitcases reliably – because it doesn't know the social world of trophies and suitcases and, at the time that example was invented, that social world was not represented on the internet.⁸ Given this, one can see that one ambitious aim of AI might be to reproduce a brain that can grasp the collective understandings of any human group into which it is conversationally embedded while a still more ambitious aim might be to reproduce a whole group of human-like AIs, with its own creatively developed collective understandings, not necessarily familiar to existing humans: that's two of the six 'levels' of AI – levels III and level V – shown in Table 1 and set out in more detail below.

Level I: engineered intelligence

The first level is the engineered intelligence which we already live with. Bear in mind that some people think that a simple thermostat is intelligent.⁹ Engineered intelligences control not only washing machines but power stations

⁸The sexism and racism of computers that strip language from the internet is not a failure but a triumph for deep learning: it shows it has a capacity for socialisation even if what is being acquired is a sad reflection of the embedding society. And, remember, it is easy to make a computer cope with any specific example once the specific problem has been pointed out and even in the act of describing the trophy/suitcase problem the potency of the internet in respect of that example is being enhanced. But the trophy/suitcase problem stands for an indefinite number of new examples which can be invented or occur naturally as society and language changes.

⁹Russell and Norvig, 2003, pps 48–52.

and missile launchers, so they have the potential to destroy us by accident without using much in the way of intelligence at all; that's not a criticism of AI, it is just a way of separating the danger of computers from their intelligence. Mostly engineered intelligence is a good, life-enhancing, thing that we would hate to be without now we have such a lot of it.

Level II: asymmetrical prostheses

A 'prosthesis', as the term is being used here in the context of AI, is something that fits into society and does the job that a human once did. We might take a calculator as an example – it does arithmetic instead of a human – but so are many of the examples that populate the previous level. So, the only difference between this level and the previous one is the extent to which people *think* that what is going on is real human-like intelligence; what humans think is rarely a matter of sharp distinctions. We can see that a thermostat is at the lower end of Level I and we can see that something like Siri, or Alexa, is at the top end of Level II, and where Level I ends and Level II begins is not that important. (In contrast, as we will see shortly, where Level II ends and Level III begins is hugely important.)

Even though the boundary between the first two levels is fuzzy, it is useful to have a Level II because it makes what is meant by 'asymmetrical' a little clearer. A crucial feature of human interaction with other humans, and with machines and other material objects, is what is called 'repair'. Starting with humans, when I am muttering indistinctly to you, you will mostly manage to work out what I am saying from the context without having to ask me to clarify. That's repair: I speak in some kind of broken or incomplete way and you use the context to fill in the gaps and file off the sharp edges to make a smooth and well-formed piece of communication. Without that, every act of communication would have to be perfect, or we would have to be continually repeating ourselves, and this would make communication very cumbersome.

This human talent for context sensitive repair is also continually deployed in our interaction with machines. For example, when we think, 'my calculator is a lot better at arithmetic than I am', it is because we continually repair its mistakes without noticing. To give an example from my 1990 book, if I want to know my height in centimetres, given that my height in inches is 69, and there are 2.54 cm to the inch, and I key 69×2.54 into the calculator, it returns, '175.26' in an instant – better than I could do in an instant – but 175.26 is not my height (at least, not for more than a fraction of a second between breaths and depending on the state of my hair), but I unthinkingly repair it to 175 cm. So in that sense the calculator is not as good at arithmetic as me because it does not know how to understand social context in a way that would cause it to approximate appropriately in the context of discussion of human height – part of the skill of good arithmetic. The calculator does not

understand social context of arithmetical calculations of human tallness in the same way as deep learning translators do not understand the social world of trophies and suitcases.

Now, this is important, because AI ‘boosters’, not to mention various philosophers, psychologists and sociologists, think that anything that enters our social life and has an effect on it as a prosthesis – and calculators, word processors, Siri and Alexa, certainly do have such an effect – should be treated as social creatures. They want them to be treated as nodes in the networks of relations that describe our social lives that are indistinguishable from other humans or, even as full-scale social intelligences.¹⁰ But they are not full-scale intelligences because if you *talk to them* in context-dependent and otherwise damaged ways they won’t be able to repair *your* output in a satisfactory way. Predictive text and spell-checkers do their best at repair, indicating that their developers know there is a problem, but they are clunky toys rather than serious contenders.

Level III: symmetrical culture-consumers

So that is why it is useful to have a Level II category of asymmetrical prostheses even though it is not clearly distinct from Level I. Level II, to repeat, turns on prostheses, the output of which we find effortlessly useful because we automatically repair that output without noticing it, just as we do with other humans. It is useful so that we can contrast it with the category of *symmetrical* prosthesis, which can effortlessly repair *our* output with as much context-sensitivity as we can repair theirs; this is Level III. At the moment, Level II AI is continually being confused with Level III and trumpeted as being real artificial intelligence, passing the Turing Test, and so on. But the jump from Level II to Level III is huge and so is the jump from where we are to serious artificial intelligence: as can be seen, it is going to involve the AI’s being effortlessly embedded into society so they can understand social context as well as humans understand social context; when they can do that, they will be able to pass properly designed, demanding, Turing Tests, rather than demonstrate a facility with games. When they can do that, they will be able to absorb human culture in the way that humans do. That is why, in Table 1, they have been given the label, ‘symmetrical *culture-consumers*’, rather than prostheses.

Impact of AI on our understanding of knowledge

In the Introduction I said I thought that the science of AI had already taught us a lot about knowledge and intelligence. So much has it taught us that I think the

¹⁰For those who know the field of STS, Latour’s so-called ‘actor network theory’ is the most popular and notorious example of this kind of elision but it is a very widespread mistake even among those who do not choose to elevate it to the central plank of a theory of the world.

philosophy and psychology of skill and expertise must change to take account of it; we now have to talk of ‘knowledge’ not ‘human knowledge’ if we want to understand any kind of knowledge including human knowledge. Before AI came along, the philosophy of skill and expertise had to do only with what humans could do; to study skill and expertise was to study humans. But now we need to change the focus away from humans and to the knowledge stuff itself. To understand knowledge, we must understand what machines can and cannot do just as much as what humans can and cannot do. For example, one huge change wrought by AI research, albeit inadvertently, is to our understanding of what counts as the apogee of knowledge. Before AI the apogee was taken to be somewhere around the things that humans find really difficult and high accomplishment in which was lauded and rewarded: when I was a school, this was the ability to do mental arithmetic and in adults it was the ability to do a really tough integration or some other such mathematical *tour de force*. But now we can do that kind of arithmetic with a pocket-sized calculator and the programme *Mathematica* can manage the fancier stuff, so that kind of thing is no longer seen as the apogee. The apogee is now seen as somewhere around some of the things that humans find easy and are still beyond computers – such as fluent language speaking.¹¹ Deep learning’s huge success in improving language handling is therefore very impressive but its failure to handle such things as trophies and suitcases shows how far there is to go: we are still at Level II rather than to Level III even though the Level II accomplishments are nuzzling the boundary.

A good way to see the difference is to consider the accomplishments of AlphaZero, which taught itself to be world champion at Go and at Chess in a matter of days, again, accomplishments that were thought to be the apogee of human accomplishments until very recently and may still not have fallen in most peoples’ estimations given the short time that has elapsed since these peaks were conquered. But both Chess and Go, even though the perfect game cannot be calculated through to the end, in the way that Noughts and Crosses can, are still played in a fixed format according to fixed rules with a fixed end-point. To reveal Level III ability, when you sit down to play Chess or Go against AlphaZero in April 2020, it would have to do what humans do and make a bit of small talk about how things were going for you in the Coronavirus crisis, before it even thought of making a move. It would have to know what fluent social interaction comprised in the current context. AlphaZero is still somewhere in the first two Levels.¹²

¹¹Except in the films where all the intelligent computers and robots are effortlessly fluent language understanders and speakers even while they are psychopaths: Hector Levesque’s (2017) definition of AI, ‘the study of how to make computers behave the way they do in the movies’, has it about right.

¹²As Shunryu Colin Garvey pointed out to me, there is a well know critical remark from the 1970s to the effect that an ideal AI of the time could make a perfect chess move while the room was on fire. There are many examples of what Level III AI’s need to be able to do in the way of editing text but Level II AI’s can’t do in Chapter 10 of *Artificial Intelligence*.

Who should be criticizing AI?

The complaint from AI enthusiasts will be along the lines that ‘every time we accomplish something new and magnificent, we’re told by the critics that if it can be accomplished it can’t be the real thing after all.’ And something has gone wrong if the critics are doing this continually – it they are continually making the goal of human-like AI an ever-receding target. But the thing that has gone wrong isn’t to be found in the critics’ domain; the thing that has gone wrong is that it should be the AI enthusiasts who are getting in first and pointing out what they have not yet accomplished in spite of the fancy and unexpected results, even when those results reach a target that the critics said could not be reached. That’s how other sciences go: in those sciences the aim is exactly defined and the worst sin is to make the claim that the aim has been accomplished and have it turn out not to be so.¹³ Maybe, as some of them claim, the deep learning community will get there through a huge increase in artificial brain capacity, maybe not.

Level III to Level IV

Notice that the aim of AI in which I am interested is learning to understand human-like intelligence by trying to simulate it with non-human means. The criterion that I am taking to indicate the achievement of human-like intelligence is passing *demanding* versions of the Turing Test – ‘DTT’s that demonstrate a grasp of social context and the corresponding ability to repair broken speech in a human-like way. Now, it seems to me that, currently, the most promising route to this goal will include the building of machines that mimic the mechanisms of the human brain in some abstract sense; I am impressed by the argument that what we need to build are better and better versions of hierarchical pattern recognizers. But that’s not where my expertise lies so my hunch in this respect is not worth much. It may be that the internal mechanism of the artificial entity that meets my criterion will be different to that of the human brain. Ava, the AI imagined in the film, *Ex Machina*, Samantha, from the film *Her*, and HAL, as portrayed in *2001, A Space Odyssey*, are thoroughly context-sensitive, fluent English speakers (who just happen to be psychopaths). In at least two of the cases, Samantha and HAL, they are imagined as being dissimilar to humans in terms of their physical construction. Nevertheless, they are still Level III devices and meet my criterion of mimicking human-like

¹³And to be crystal clear, when AI reaches Level III, and according to at least some people who do understand what a demanding Turing Test would look like, they think this will be accomplished pretty soon, perhaps even in my lifetime if I am lucky, I will be delighted and ready to say that human intelligence has been simulated artificially. (What a demanding Turing Test would look like is explained in Chapter 10 of my 2018 book.) That is, I will be delighted provided the AI enthusiasts have not surreptitiously shifted the goalposts themselves by aiming to convince an audience with some tricks rather than truly aiming for linguistic fluency in the face of the hardest, context dependent, tests. Sadly, this kind of goalpost shifting also happens all the time.

intelligence. Such devices might be a little disappointing in that we may not learn as much about human intelligence from mimicking it in this way as we would have learned from mimicking the human mechanism, but maybe in recognizing we are doing something different we will still be learning what the real mechanism is.¹⁴ The point is, that there is also a Level IV where the mechanism is broadly the same.

Level IV of artificial intelligence: humanity-challenging culture-consumers

The difference between level III and level IV is a subtle one and hard to pin down exactly. At Level III AI achieves human-like intelligence but at Level IV the mechanism by which it achieves it must be the same as when it is exhibited by humans. The question of similarity of process will remain important for those interested in AI as a route to proving that humans are just meat machines without free will, and that humanity doesn't have a soul nor anything unique that could stand in for it. It will also remain important for those who think AI is the route to understanding human intelligence even if they are not pre-committed to any metaphysical view.

Unfortunately, whenever the notion of 'the same' is deployed there is always going to be a problem about what it means. For example, would a Level IV device need human-like consciousness (which is not a precondition for Level III)? Since humans themselves can carry out the same actions with varying degrees of consciousness (e.g. it is claimed that a mark of truly skilled physical performance is an absence of conscious attention), it is hard to foresee where the argument about the need for consciousness will go in the case of Level IV. If we are not sure if consciousness is a precondition for intelligent action in humans how can we claim it is a precondition for achieving intelligence in the same way as humans achieve it?

One must not make the answer to what constitutes the human process a truism by insisting that doing things like humans means using the same *biological* mechanisms or the problem becomes not one of reproducing *human intelligence* but reproducing *humans*. So, we must accept that thinking 'like' a human while using silicon chips or some such, potentially meets the criterion of Level IV – reproducing human internal states. But what about AlphaZero; setting aside the small-talk problem, you could not play a decent game of Chess or Go with it because it would win every time. It could be claimed that it is still using the human thought processes that humans use when they play games – hierarchical pattern recognition – but just doing it much better! So, nice philosophical questions remain at this transition point.

¹⁴The idea of interactional expertise is important here. Some argue that a human-like body is necessary to achieve human-like understanding sufficient to pass a Turing Test. The question is discussed in, for example, Collins, 2020, but this argument probably has some way to go.

Levels V and VI

Level V is like Level III, or Level IV, except that the AI's will be sufficiently human-like so that groups of them can develop human-like cultures by themselves. Here the question of the necessity of a human-like body, which is disputed at Levels III and IV, is resolved. For a group of machines to develop human-like cultures they will need human-like bodies because, while it can be argued individuals can *acquire* human-like culture through immersion in language alone, without participating in the physical activities of a culture (interactional expertise), we can't *develop* human-like cultures, nor sustain them over the generations; we must have a body-type which affords the corresponding physical activities.¹⁵ So such machines will need bodies that could play tennis and cricket and American football and snooker, and so on, at about the same level as humans, even if they don't wind up playing them but develop their own sports or reject sport altogether. Solutions to the problems of human-like robotics are going to be essential at Level V. Thus, while an autonomous society of intelligent dogs – dogs with a more elaborate speaking apparatus – might develop a new language, it couldn't include words for tennis racket or cricket bat, at least, not sustainably, unless the dogs encountered humans and maintained linguistic contact with them.

Humans continually try to develop new cultures. Sometimes these turn on physical appearance. Thus, there is currently a half-jokey movement under way to bring out the cultural specificity of red-haired people. So, let us equip our otherwise human-like AIs with a metallic-looking silver skin so they can be easily identified by others and easily recognize each other. We can then imagine them forming their own cultural group, proudly or even aggressively distinct from the other human cultures around them and pulling away from existing human societies (this, of course, is also the stuff of science fiction). That would be Level V of artificial intelligence.

This seems to be what those fearful of the 'singularity' are thinking of when they claim that the computers will one day be 'so intelligent' that we will be lucky if they are willing to keep us as pets. The doom-mongers see intelligence as a monotonic accomplishment, of which you simply have either more or less, and if you have more you automatically become more powerful and dangerous. But there has to be something special about the intelligence if it is going to be inclined to overpower the humans who made it; it has to be the kind of intelligence that is capable of forming its own cultures. That culture may not be a violent one; perhaps it will be a peaceful culture – there are many such. But it may pick up its cues from the violent intentions in human societies. It is probably sensible not to try to make Level V just in case it does develop in a way we will regret.

¹⁵For interactional expertise see Collins and Evans 2015 and Collins, 2020

Level VI departs from the AI aspiration set out at the beginning of the piece: ‘attempt to explicate the workings of human intelligence by creating “mechanical causes and effects that mimic human actions”.’ Level VI may still be trying to explicate the workings of human intelligence but, if it is, it will be doing it by creating non-human kinds of intelligence. Taking our cue from science fiction, once more, it might try to create an artificial version of the intelligence of the extra-terrestrial heptapods portrayed in the film *Arrival*. It will be difficult to know whether we have created such an intelligence because we would not know what questions to ask in a demanding Turing Test; it would be like a Turing Test to distinguish between a machine mimicking a Chinese-speaker and a human Chinese-speaker but where the judge does not speak Chinese. I don’t know whether Level VI really is within AI’s project, but sciences do develop their own momentum and it is not hard to imagine AI going this way once it has reached the other levels. Once more, it could be a hazardous undertaking.

Conclusion

What I have tried to do here is point out some features of AI and its discontents and compare them with the critical debates in other kinds of science. Compared to say, physics, the debate about AI is diffuse. I have suggested that this is because both proponents and critics of AI most often direct their arguments at outside audiences rather than inside practitioners. A stark contrast is found in the fact that in domains like high energy physics, or gravitational wave physics, it is recognized that the sharpest and most productive criticisms have to come from inside the domain whereas all too often in the world of AI insiders, hyping of products takes precedence over internally organized criticism. This leaves the field free for outsiders to generate a heterodox range of complaints, and it leaves the field vulnerable to these complaints. There is a tradition of insider criticism of AI, and, unsurprisingly, it is sharp, but the tradition seems to be getting thinner.

My criticism is based on my expertise on the nature of human understanding – that which is to be mimicked by AI – and I use it as a way of refining the aims of AI and dividing them into a possible six levels. My claim is that we are currently at the top end of Level II but still a long way from Level III. Attainment of Level III will be demonstrated by AIs that can pass suitably demanding Turing Tests (DTTs), which currently no machines can pass. The continual claims by AI boosters that the Turing Test has been passed is a problem for AI as a respectable science and as the ‘Western World’ encounters political dangers that most of us thought had long become part of history, we desperately need sciences to act respectably so they can be role models for decision-making and legitimate checks and balances or political ambitions. The six levels may function as a way of bringing some order to the ambitions

and claims about the accomplishments of AI. No doubt other ways of defining the aims of AI could be devised, though I think all of them will need to include the division between Level II and Level III. Whichever way of defining the aims gains the most widespread assent, to have a more carefully defined target should help, and that could be vitally important for far more than AI itself.

In sum, AI is one of the most important sciences in the world but its way of presenting itself is still too influenced by its early insecurities as a science. To play the role we need it to play in today's political world, a role which it can now well afford to play given its almost unprecedented financial independence as a science, it needs to curb the reflexes developed in those early days. Instead it should act like the iconic science of physics: that is, the aim should be never to announce that more has been achieved than has been achieved. This aim can be achieved only through the nurturing and honouring of internal critics rather than the rejection which was typical of the formative period. I have to add that it should be obvious by now that human intelligence is a collective enterprise with language being an iconic example: those internal critics will have to take this into account, emulating brains that don't stand alone but interact fully in society. If AI can switch to becoming this kind of strongly self-critical science, future generations would look back at it as one of the institutions that helped save pluralist democracies from populism rather than as a notorious champion of alternative facts.

Postscript: Marcus and Davis; engineering and the social

After the penultimate draft of this piece was written I came across the recently published book by Gary Marcus and Ernie Davis entitled *Rebooting AI: Building artificial intelligence we can trust* (2019), and also their critique, in *MIT Technology Review* (2020), of a recent and much-hyped language processing device known as GPT-3 (GPT = stands for Generative Pre-Training). Since both Marcus and Davis (M + D) are technical insiders in the field of AI and their work is interestingly and productively critical, and since it illustrates some of what is argued above, it seemed useful to discuss it. I referred to some of Davis's fascinating 'common-sense' criticisms and the decisive 'Winoograd schema challenge test' for existing AI, in which he was deeply involved, in my 2018 book (*Artificial Intelligence – AA*) on which my article is based. Their book is also reassuring in that these technical experts include many of the same technical elements in their discussion of AI as can be found in *AA* though, of course, they are able to include far more detail concerning technical developments.

Starting with GPT-3, it is striking that the company that makes it refused to allow M + D access to it for the purpose of testing it; they had to obtain access to it through a 'back door' (2020). Here we see, once more, AI still not acting as a respectable science encouraging technical criticism. As M + D explain,

Silicon Valley entrepreneurs often aspire to “move fast and break things”; the mantra is “Get a working product on the market before someone beats you to it; and then worry about problems later.” (2019, 188)

Once M + D were able to test GPT-3 it on questions of common sense and the like, they were able to show that it represented no significant improvement in terms of language ‘understanding’ over previous language processors in spite of the hype: it’s failures were of the trivial type discussed above (and in Chapter 10 of AA).

But their book is also revealing in the ways that it differs from mine and from the light it sheds on some of the arguments presented above. The differences arise, I believe, out of disciplinary background and approach: Marcus is a psychologist by training and Davis is a computer scientist. Furthermore, both approach the topic as an engineering problem. They state at the outset of their book:

Crucially, AI is not magic, but rather just a set of engineering techniques and algorithms, each with its own strengths and weaknesses, suitable for some problems but not for others (2019, 24)

In contrast, AA approaches the topic from the point of view of a philosophically inclined sociologist and looks at AI as a science aimed at elucidating the nature of human knowledge.

Starting with the psychology/sociology tension, though the topic of the relationship between top-down and bottom-up understanding appears in both my book and theirs, they present it as a matter of individual psychology, nicely demonstrated by the way the same individuals can be ‘primed’ with different stimuli to interpret an image in different ways. There is no discussion in their book of how whole societies, or sub-groups within societies, or adherents to different scientific paradigms, inhabit different social settings in which the world is viewed in the same way by all the inhabitants of that setting but in different ways to inhabitants of other social settings. This individual/collective contrast is a key to the difference in thinking.

Turning to the engineering versus philosophy/sociology contrast, M + D’s aim is to improve AI by preventing its being taken over by deep learning techniques which start every new task from scratch. Instead, they believe successful programmes must use older AI techniques to insert a large component of explicit physical and common-sense understandings as a foundation for any subsequent learning. Here we can see the influence of Marcus’s mentor, Steven Pinker and, in turn, his debt to Chomsky. These older techniques have not proved effective on their own but, they argue, should be much more effective when *combined* with deep learning. They also suggest inserting, ‘by hand’, some ethical principles into programmes from the start. They are engaged, then, in an internal technical battle to stop deep learning entirely taking over

the world of AI and countering this tendency with the addition of some more explicit programming.

In AA (p110 ff) I argue that to make sense of the different way different groups of humans interpret the world there must be a common foundation of perceptual abilities on which the varied interpretations are based, so I sympathize with M + D's desire to base deep learning on an explicit and universal perceptual foundation. But the big question is how deep this foundation should be.

M + D appear to want to build a deep foundation based on the knowledge of how Western societies work and including both lots of scientific understanding as well as norms of behaviour in different settings. This approach might well produce better engineered AIs for use in Western settings, but the resulting programmes are still going to be vulnerable to occasional, unpredictable, unhuman-like failures whenever they approach tasks in unrestricted or new domains and settings. Ironically, M + D are themselves experts on these kinds of problems, as Davis's excellent Winograd challenge Turing Test quoted in their book (2019, 93), in Chapter 10 of AA, and which forms the basis of their critique of GPT-3, reveals.¹⁶ They know the human world is open-ended, that the knowledge and common-sense that forms it cannot be captured in a set of formal rules and that, therefore, the addition of sets of explicit rules and facts may improve intelligent devices but only in the way that *Band Aid* improves a wound; they know this as well, or better, than anyone else. And yet they seem to forget this issue when they offer advice for making a better AI. The schizophrenia is also there when they insist that a replacement for the Turing Test is needed given that Davis's own version of a demanding Turing test is the very tool they use to show the deep inadequacy of existing AIs.¹⁷

It seems to me that the schizophrenia arises from approaching the problem as one of engineering rather than philosophy/sociology. That goal leads them to try to work out a way to build programmes that will better capture features of Western scientific culture. What they fail to notice is that different groups of humans see the world in very different ways; one cannot fail to notice this if the goal is to reproduce human intelligence. An AI that is to reproduce human intelligence will have to be itself *capable* of seeing the world in many different ways. Building AIs that start with the uniform model of the world provided by the current state of Western scientific culture will not solve the problem of human intelligence. Any built-in foundation of common perceptual abilities, such as the recognition of basic shapes and patterns and so forth has to be a shallow if it is to allow scope for all the varied perspectives of current and future human groups, if AIs, like humans, are to be able to learn different things from exposure to different social worlds.

¹⁶But see the remark about progress in respect of Winograd schemas in footnote 4 of this paper.

¹⁷I have argued at length (eg in AA) that a computer that can pass a demanding Turing Test (rather than some tricky version of it), will have solved the deep problems of AI.

M + D state:

AI that is powered by deep understanding will be the first AI that can learn the way a child does, easily, powerfully, constantly expanding its knowledge of the world. (2019, 201)

How can one disagree that ‘deep understanding’ is a likely component of human-like intelligence? How can one disagree that learning ‘the way a child does, easily, powerfully, constantly expanding its knowledge of the world’ would be a fine thing in an intelligent computer? But this is like agreeing with the virtues of motherhood and apple pie. What does ‘deep understanding’ mean and why is there no discussion in their book until page 201 of how *humans* come by their common-sense? Why is there no discussion of how you would build a computer that could occupy the social spaces that children occupy in the course of their upbringings so that they could truly learn like a child? Why is there no discussion of how this will lead to the marked variations in the substance of the intelligences of such devices when immersed in different social locations? I am suggesting that it comes from mixing up engineering solutions with understanding human knowledge. In open domains, engineering solutions are always going to be a matter of more and more *Band Aids*.

In sum, to get beyond Level 2 of artificial intelligence, and find the kind of solutions that the non-engineering side of Marcus and Davis want, which would pass their demanding Turing Tests, will require building machines that are capable, in principle, of absorbing non-Western cultures as readily as Western cultures and have the potential to absorb all the varied, cultures (including the crazy ones), found in Western societies. If a machine is to do that, it cannot be constrained by too much built-in current science and engineering so its knowledge and rules foundation will have to be a shallow one and the large preponderance of what it knows will be learned from scratch and therefore capable of being different in different settings (see also Adams; Blackwell, this issue).

When a machine has been built that can absorb all these different cultures – which means a machine has been built that can truly learn like a child – then it will be also be a machine that can absorb Western scientific culture properly rather be pre-programmed with a stick-figure caricature of Western scientific culture. It will then be able to handle the engineering problems presented by Western cultures as reliably and creatively as humans while its failures will be human-like. Ironically, then, really good engineering solutions will need to solve the philosophical and sociological problems first!

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Harry Collins is Distinguished Research Professor at Cardiff University. He is an elected Fellow of the British Academy and winner of the Bernal prize for social studies of science. His 25 books cover, among other things, sociology of scientific knowledge, artificial intelligence, the nature of expertise, tacit knowledge, and technology in sport.

References

- Collins, Harry. 2010. *Tacit and Explicit Knowledge*. Chicago: University of Chicago Press.
- Collins, Harry. 2017. *Gravity's Kiss: The Detection of Gravitational Waves*. Cambridge, MA: MIT Press.
- Collins, Harry. 2018. *Artificial Intelligence: Against Humanity's Surrender to Computers*. Cambridge: Polity Press.
- Collins, Harry. 2020. "Interactional Imogen: Language, Practice and the Body." *Phenomenology and the Cognitive Sciences* 19 (5): 933–960. doi:10.1007/s11097-020-09679-x.
- Collins, Harry, Robert Evans, Darrin Durant, and Martin Weinel. 2019. *Experts and the Will of the People: Society, Populism and Science*. Basingstoke: Palgrave.
- Collins, Harry, and Evans Robert. 2015. "Expertise Revisited I - Interactional Expertise." *Studies in History and Philosophy of Science* 54: 113–123. (a pre-print is available at <http://arxiv.org/abs/1611.04423>).
- Collins, Harry, and Evans Robert. 2017. *Why Democracies Need Science*. Cambridge: Polity Press.
- Levesque, Hector. 2017. *Common Sense, the Turing Test, and the Quest for Real AI*. Cambridge, MA: MIT Press.
- Levesque, Hector, Ernest Davis, and Leora Morgenstern. 2012. "The Winograd Schema Challenge." *Proceedings of Principles of Knowledge Representation and Reasoning*.
- Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Vintage.
- Marcus, Gary, and Ernest Davis. 2020. 'GPT-3, Bloviator: OpenAI's Language Generator has No Idea What It's Talking about. *MIT Technology Review*, August 22.
- Russell, Stuart J., and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*. 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies

Inga Ulnicane , Damian Okaibedi Eke , William Knight , George Ogoh 
and Bernd Carsten Stahl 

Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK

ABSTRACT

Recent advances in Artificial Intelligence (AI) have led to intense debates about benefits and concerns associated with this powerful technology. These concerns and debates have similarities with developments in other emerging technologies characterized by prominent impacts and uncertainties. Against this background, this paper asks, What can AI governance, policy and ethics learn from other emerging technologies to address concerns and ensure that AI develops in a socially beneficial way? From recent literature on governance, policy and ethics of emerging technologies, six lessons are derived focusing on inclusive governance with balanced and transparent involvement of government, civil society and private sector; diverse roles of the state including mitigating risks, enabling public participation and mediating diverse interests; objectives of technology development prioritizing societal benefits; international collaboration supported by science diplomacy, as well as learning from computing ethics and Responsible Innovation.

KEYWORDS

Artificial Intelligence;
emerging technologies;
governance; policy; ethics;
regulation; societal
challenges; Responsible
Innovation

Uncertainty about the impact of AI can be a concern but it is also an opportunity: the future is not yet written. We can, and should, shape it. (European Commission 2018a, 13)

Introduction

Is time travel among the numerous wonders promised by Artificial Intelligence (AI)? Could this transformative and revolutionary technology transport us back to ‘good (or not so good) old times’? For researchers of governance, policy and ethics of emerging technologies, recent years of academic and public debates about AI have often presented an opportunity to travel several decades back

CONTACT Inga Ulnicane  inga.ulnicane@dmu.ac.uk  De Montfort University, The Gateway, Leicester LE1 9BH, UK

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

to the twentieth century, namely, to the times when emerging technologies were understood mostly in terms of their contributions to economic growth and national prestige. Actors involved in innovation were largely limited to industry and academia. The state mostly was seen as having a limited role of market correction and the field of computing ethics was less prominent. Those were times with little or no awareness of cross-cutting political, social and ethical issues of emerging technologies – from life sciences to information technologies – that shape our lives (e.g. Jasanoff 2016; Juma 2016) and specific ways to govern them (e.g. Kuhlmann, Stegmaier, and Konrad 2019) such as Responsible Innovation (e.g. Stilgoe, Owen, and Macnaghten 2013). The knowledge accumulated and lessons learned about governance, policy and ethics of emerging technologies over the past decades have not much featured in recent discussions about AI.

Is AI really such a *sui generis* phenomenon that its governance and ethics have little if anything to learn from other emerging technologies? A number of concerns and issues addressed in AI debates leave a *déjà vu* feeling because they are well-known from work on other emerging technologies. This suggests that when addressing concerns associated with AI, rather than ‘reinventing the wheel’, there are opportunities to learn from advances made and lessons derived from governance, policies and ethics of other emerging technologies. While each technology has some unique features, all emerging technologies have a number of common characteristics – such as radical novelty, relatively fast growth, coherence, prominent impact, and uncertainty and ambiguity (Rotolo, Hicks, and Martin 2015) – that pose some similar concerns and allow to learn across different technologies.

Against this background, this paper asks – what can AI governance, policy and ethics learn from other emerging technologies to address concerns and ensure that AI develops in a socially beneficial way? To answer this question, we draw on analysis of AI policy documents and on recent literature on governance, policy and ethics of emerging technologies. Discussions of concerns about AI and governance, policy and ethics of emerging technologies are diverse and extensive. It is beyond the scope of this article to cover them comprehensively and to study all of them in-depth. Therefore, the main focus here is on some of the key concerns and solutions identified in AI policy discussions and lessons from work on other emerging technologies that might be relevant for AI.

With this article, we aim to contribute to the social studies of AI and emerging technologies more broadly with a particular focus on their governance and policy. Our examination of recent policies for AI is complementary to other articles in this special issue on AI discontents (Garvey 2021, this issue) engaging, for example, with critical analysis of AI for good initiatives (Holzmeyer 2021, this issue), emerging AI applications in healthcare (Datta Burton et al. 2021, this issue) and social criticism of computing (Loeb 2021, this issue). While our article takes a ‘bird’s eye’ view on common trends in recent policies

for AI, similar to other contributions in this special issue (Adams 2021; Blackwell 2021, this issue) we recognize the importance of diverse national, regional and local contexts and cultures.

This article proceeds as follows: first, insights from AI policy documents on AI framing, a mix of hopes and concerns and suggested solutions in terms of ethics and regulation are discussed; second, we reflect on the lessons that recent literature on governance, policy and ethics of emerging technologies can offer to AI; and finally, we conclude summarizing the suggestions from this literature that could be relevant for AI governance.

AI policy debates: framing of hopes, concerns and solutions

Recent advances in AI, driven by developments in hardware and big data (see, e.g. Marcus and Davis 2019), have triggered active public debates around the world about the benefits and concerns related to AI and appropriate public policies (Ulnicane et al. forthcoming). According to the OECD, in early 2020 at least 50 countries¹ have developed, or are in the process of developing, national AI strategies.² Additionally, international organizations, consultancies and stakeholders have also launched their AI policy documents. We have analysed 49 AI policy documents launched from 2016 to 2018 by national governments, international organizations, consultancies and stakeholder organizations in Europe and the United States (for the list of documents, see the Appendix).³ In our analysis, we focused on how these policy documents frame AI, associated benefits and concerns, and what mechanisms they suggest for addressing them.

In these policy documents, we find various definitions and understanding of AI as well as concerns about difficulties to define AI (The 2015 panel 2016; Villani 2018; European Commission 2018a; EESC 2017). While it is recognized that the concept of AI has existed for more than 60 years, these documents highlight that the real-world applications have only accelerated during the past decade (Campolo et al. 2017; Crawford and Whittaker 2016; European Commission 2018a Executive Office of the President 2016a; UNI Global Union 2017). Wide-ranging and long-lasting effects of AI across many areas of life and numerous sectors of economy are often discussed. Sometimes they are presented as a reason why AI is different from other technologies, as can be seen in a French document which states that

¹Most of these AI strategies have been launched in Europe, North America and major Asian economies. Similar geographical concentration can be seen with regard to AI ethics guidelines (Jobin et al 2019).

²AI strategies and public sector components <https://oecd-opsi.org/projects/ai/strategies/> Last accessed 15 February 2020.

³We selected these policy documents according to a set of criteria: they have strong focus on AI; they focus on overarching AI policy rather than AI policy in a specific domain such as education and health; and they address policy questions rather than just ethical principles. For more information on methodology, see Ulnicane et al 2020.

The key factor setting AI apart from other scientific disciplines is its all-encompassing impact society wide. This is not just some passing trend or media phenomenon, far from it: its implications are posed to be long-lasting and game-changing worldwide. AI is seeping into all sectors – economic, social, political and cultural alike [...] The key question now is nothing less than what kind of society we wish to live tomorrow. (Villani 2018, 63)

Attitudes towards these numerous long-term effects of AI represent a complex mix of optimism on how they should improve ‘our lives’ (Villani 2018) and caution or even concern pointing out that ‘it is necessary to look carefully at the ways in which these technologies are being applied now, whom they’re benefitting, and how they are structuring our social, economic, and interpersonal lives’ (Crawford and Whittaker 2016). Some documents more emphasize expected benefits of AI, others focus more on concerns and yet other ones attempt to balance opportunities and challenges associated with AI. A common feature is that, typically, the impact of AI is seen to be significant across many areas, from jobs, health and education to transport and security. While discussing opportunities, some documents predominantly focus on economic impact in terms of increases in growth and productivity, while others take a broader view considering societal benefits. Latter ones include hope that AI ‘will be central to the achievement of the Sustainable Development Goals (SDGs) and could help solve humanity’s grand challenges’ (ITU 2017).

In context of this paper’s focus on emerging technologies, it is interesting that one document explicitly indicates that benefits are related to AI interaction with other technologies, stating that ‘in combination with other emerging and converging technologies, AI has the potential to transform our society through better decision-making and improvements to the human condition’ (Bowser et al. 2017). This idea sees AI as a critical component of the so-called fourth industrial revolution that includes the fusion of physical, digital and biological technologies (Schwab 2017). Some documents simplify or selectively use examples from previous emerging technologies, for example, by suggesting that ‘the history of technology development tends to show that a foundational technology, such as the Internet, can serve everyone’ (Accenture 2017) and comparing ‘the advent of AI technologies to the development of the commercial internet in the 1990s to provide insight into how policy-makers may champion pro-growth policies while maintaining an appropriate level of oversight and accountability for consumers’ (Thierer, Castillo O’Sullivan, and Russell 2017). These two examples of internet history ignore more problematic issues such as digital divide or surveillance.

Concerns and challenges

The policy documents address a wide variety of concerns and challenges associated with AI including ethics, safety, privacy, transparency and accountability,

work, education and skills, inequality and inclusiveness, law and regulations, human rights and fundamental norms and values, governance and democracy, and warfare (e.g. EESC 2017). Some concerns are phrased more as individual level ethical questions about personal integrity, autonomy, dignity and freedom of choice (e.g. EESC 2017), while others address macro-level issues of geopolitics, power and populist political movements (Campolo et al. 2017). Some of the key concerns are summarized in this quote from a European Union document:

Workers fear they will lose their job because of automation, consumers wonder who is responsible in case a wrong decision is taken by an AI-based system, small companies do not know how to apply AI to their business, AI startups do not find the resources and talent they need in Europe, and international competition is fiercer than ever with massive investments in the US and China. (European Commission 2018b, 1)

Some policy documents state that challenges posed by AI are similar to those raised by other emerging technologies but others argue that such comparisons are not relevant for AI due to its major differences from earlier technologies. A US document states that ‘as with most transformative technologies, AI presents some risks in several areas, from jobs and the economy to safety, ethical, and legal questions’ (Executive Office of the President 2016a). Similar ideas can be found in the European stakeholder document which tells that ‘as with every disruptive technology, AI also entails risks and complex policy challenges in areas such as safety and monitoring, socio-economic aspects, ethics and privacy, reliability, etc.’ (EESC 2017) However, a document from the European Parliament suggests that AI is very different from previous technologies highlighting that ‘the split between past and future societal models will be such that we cannot expect to take the emergence of information technology, the internet or mobile phones as a starting point for reflection’ (European Parliament 2016). This quote from a report from the UK Parliament summarizes some old and new concerns posed by AI:

Many of the challenges now linked to AI are far from new. For instance, concerns about increasing inequality gaps, stereotypes and biases, shortages of skills, and abuse of power have existed in our society for centuries now. AI is not the creator of these problems. Rather, in many ways, AI is simply resurfacing prevailing problems and urging society to acknowledge their existence and provide solutions. On the other hand, AI technologies are of such high impact and progress at such rapid speeds that some issues developing are authentically new. Some of these include increasingly automated decision-making, potentially catastrophic security threats, technological unemployment, and transformations in current notions of privacy, agency, consent, and accountability. (BIC/APPGAI 2017a, 6)

Proposed solutions: ethics and regulation

How can these old and new concerns raised or reinforced by AI be addressed? Almost every AI policy document calls for an appropriate ethical and legal framework and related activities such as ethics education and research on ethical,

legal and social implications of AI. A US document points out that this call has similarities and differences with other technologies: ‘as with any technology, the acceptable uses of AI will be informed by the tenets of law and ethics; the challenge is how to apply those tenets to this new technology, particularly those involving autonomy, agency and control’ (Executive Office of the President 2016a). Due to the global reach of AI, several documents suggest that ethics guidelines and sometimes even regulation should be coordinated or adopted at international level (e.g. EGE 2018; Rathenau Institute 2017).

Interestingly, while policy documents often mention AI ethics and law next to each other, suggesting that they are closely linked, on a closer reading and observation of practical developments, the major differences in attitudes to ethics and law emerge. The documents reveal a lot of enthusiasm for ethics codes and guidelines based on human rights and values that would guide the development and use of AI. Only occasionally more cautious notes are mentioned reminding that AI ethics codes ‘should be accompanied by strong oversight and accountability mechanisms’, that ‘more work is needed on how to substantively connect high level ethical principles and guidelines for best practices to everyday development processes’ and the ‘need to move beyond individual responsibility to hold powerful industrial, governmental and military interests accountable’ (Campolo et al. 2017).

If the overall attitude towards ethics guidelines is enthusiastic, then views on regulation are typically more cautious with caveats added. Often it is suggested that while regulation is needed to avoid AI related risks, it is important to ‘avoid the risk of over-regulation, as this would critically hamper’ innovation (European Commission 2017) and that ‘regulation that stifles innovation, or relocates it to other jurisdictions’ would be counterproductive (The 2015 panel 2016). Some caveats added to regulation discussion include pointing out that there is a lot of uncertainty about this new technology and ‘taking the right approach to laws and regulations on AI will also require a good understanding of what AI can, cannot and will be able to do in the short, medium and long term’ (EESC 2017) and that ‘attempts to regulate “AI” in general would be misguided, since there is no clear definition of AI (it isn’t any one thing), and the risks and considerations are very different in different domains’ (The 2015 panel 2016). This cautious approach includes statements that the officials are monitoring the developments, and reviewing and adapting existing legal frameworks (European Commission 2018c). Interestingly, these caveats and cautious statements are hardly ever mentioned when discussing codes of ethics, to which they might also be relevant.

While ethics guidelines and regulations are interconnected as regulations can be a way to protect values and implement voluntary ethical guidelines via binding legislation, there are important political, technical and other types of differences between the two, which is part of the explanation why more has been done in launching AI ethics guidelines than adopting regulation. A recent review of 84 AI ethics guidelines adopted by governments, international

organizations, think tanks, companies and professional organizations (Jobin et al 2019) confirm vibrant developments in this field. Much less activity has taken place in adopting regulation. For example, the European Union has been much faster in launching its AI ethics guidelines, while discussions on appropriate regulation and legislation take much longer (Ulnicane 2021; Vesnic-Alujevic, Nascimento, and Polvora 2020).

Experts have suggested that active work on ethics and little progress on regulation are connected because:

AI ethics initiatives have thus far largely produced vague, high-level principles and value statements that promise to be action-guiding, but in practice provide few specific recommendations and fail to address fundamental normative and political tensions embedded in key concepts (for example, fairness, privacy). Declarations by AI companies and developers committing themselves to high-level ethical principles and self-regulatory codes nonetheless provide policymakers with a reason not to pursue new regulation. (Mittelstadt 2019, 501)

This idea that companies strategically promote ethics as part of their public relations to delay or avoid binding regulation is echoed by others. For example, Thilo Hagendorff (2020) points out that the focus of companies on ethics, which lacks mechanisms to enforce its normative claims, discourages efforts to create a binding legal framework. A number of stories from Europe and the US (Coeckelbergh and Metzinger 2020; Metzinger 2019; Ochigame 2019) provide empirical insights that support the above claims.

If regulation tends to be avoided or delayed and the role of ethics is contested, then what remains to address the concerns posed by AI? Below we discuss a number of insights from other emerging technologies on their governance, policy and ethics that can provide suggestions for addressing AI related concerns.

Towards a ‘good governance’ of AI? Lessons from governance, policy and ethics of emerging technologies

Recent years and decades have seen the emergence and development of ideas, concepts and practices for shaping the development and use of emerging technologies towards societal benefits. In particular, work on the inclusion of diverse stakeholders, consideration of various roles of the state, broad societal goals of technology, international cooperation and science diplomacy, computing ethics and Responsible Innovation are discussed here. While each of these ideas, concepts and practices also have some limitations, they can provide useful starting points to broaden and enrich approaches to AI, which faces a number of challenges similar to those for other technologies. These are some opportunities to shape AI development and use nationally and internationally as well as at the level of research projects and laboratories, which need to be further contextualized to specific locations, cultures, temporalities and applications.

Beyond government: governance

AI policy documents (e.g. BIC/APPGAI 2017a) tend to mention governance as a way to facilitate benefits and mitigate risks associated with AI (Ulnicane et al 2020). However, AI documents (and the academic literature) hardly ever define or explain what do they mean by governance. Either governance is mentioned close to government implying that governance is something that government does or it is mentioned next to ethics suggesting that governance is similar to and reinforces ethics. Looking at the social science literature on governance and its role in emerging technologies can help not only to clarify definitions but also to reveal the potential of systematically addressing governance issues for AI.

In the social science literature, a move from government to governance (Pierre and Peters 2000) stands for the involvement of more diverse non-governmental actors, groups and networks from civil society and private sector in decision-making and coordination. According to Susana Borrás and Jakob Edler (2014, 13–14), governance is understood ‘as the mechanisms whereby societal actors and state actors interact and coordinate to regulate issues of societal concern’. Coordination including a broad range of stakeholders is of high relevance for emerging technologies. Specific governance approaches have been suggested to address the uncertainty surrounding future developments, societal benefits and risks associated with emerging technologies. For example, the notion of ‘tentative governance’ (Kuhlmann, Stegmaier, and Konrad 2019) emphasizes the role of flexibility, learning and reflexivity in governing emerging technologies.

Elements of governance which emphasize the importance of interaction between government and a broad range of societal actors in decision-making can be found in AI policy documents which suggest multi-stakeholder approaches, inclusion and dialogue to ensure that AI is developed and used according to interests and needs of the society (Ulnicane et al 2020). How these important ideas are implemented is crucial. Some early experiences with multi-stakeholder AI forums (e.g. Metzinger 2019) suggest that they face the risks well-known in social science literature on the collective action, namely the problem that arises when business interests are better organized and resourced, and forums for public participation can be captured by vested interests (Olson 1974). It is important to be aware of such risks and take actions to ensure balanced participation of diverse interests.

Beyond market correction: diverse roles of the state

What is the role of the state in the governance of emerging technologies? This is a highly relevant question for AI because of concerns about the dominant role of big technology companies in this field in which ‘the vast majority of the development of AI and all its associated elements (development platforms, data, knowledge and expertise) is in the hands of the “big five” technology

companies' (EESC 2017). If power and resources in this field are concentrated in a small number of large companies, then the question arises what can the state do to influence its development in societal interests? Policy documents suggest a number of activities expected from the state including regulation and supporting research and retraining. For example, a French document assigns the state the role of a key driver for 'laying the foundation for innovation and providing stakeholders with the means and the resources for breaking new ground' (Villani 2018). The AI Now 2017 report with a particular focus on the US situation highlights that the role has shifted during the history of AI:

In the mid-twentieth century, advanced computing projects tended to be closely associated with the state, and especially the military agencies who funded their fundamental research and development. Although AI emerged from this context, its present is characterized by a more collaborative approach between state agencies and private corporations engaged in AI research and development. (Campolo et al. 2017, 23)

There are important questions to be asked about this collaborative approach between the state and private companies not only in the context of the US administration but more generally: Whose interests do they serve? Do they consider broader societal interests or are they predominantly geared towards economic interests of companies? Are other groups of civil society, consumers and users included in these collaborative relationships? Are these relationships transparent and accountable to the public?

One way to approach these questions is to reflect on the role of the state in the field of emerging technologies. If traditionally the role of the state with respect to new technologies was associated with market correction, then recent work in the innovation policy studies has undertaken a more systematic approach to identify diverse roles that the state plays. Borrás and Edler (2020) have identified 13 roles of the state and many of them are highly relevant for addressing some of the AI related concerns. This includes such roles as, for example, a *mitigator* trying actively to reduce the negative effects that arise as a consequence of socio-technical change; an *enabler* of societal engagement encouraging the involvement of stakeholders in participatory processes to define direction of change; and a *moderator* acting as arbitrator or negotiator between different social and political positions among agents regarding the direction of transformation of a socio-technical system (Borrás and Edler 2020). Considering the variety of roles that the state can play allows to think more broadly about the opportunities for the state to shape the development and use of AI.

Beyond growth: policy to address societal challenges

What is the objective of AI policy? Many AI policy documents still follow traditional paradigm of technology policy focusing on economic contribution of AI to increase growth and productivity. However, some documents also

include more recent shift in technology policy towards societal objectives mentioning potential contribution of AI to addressing important social challenges, achieving the United Nations Sustainable Development Goals (ITU 2017; Vinnova 2018) and the European Green Deal tackling climate and environmental-related challenges (European Commission 2020). In policy documents, traditional and novel ideas about the objectives of technology policy occasionally are mixed, as it can be seen in this quote from this French document which talks about the paradigm shift towards energy-efficiency but at the same time sticks to traditional discourse of growth and economics:

A truly ambitious vision for AI should therefore go beyond mere rhetoric concerning the efficient use of resources; it needs to incorporate a paradigm shift toward a more energy-efficient collective growth which requires an understanding of the dynamics of the ecosystems for which this will be a key tool. We should take the opportunity to think of new uses for AI in terms of sharing and collaboration that will allow us to come up with more frugal models for technology and economics. (Villani 2018, 102)

The traditional technology policy paradigm focusing on economic growth and productivity is increasingly challenged in the context of climate change and resource consumption (see e.g. De Saille et al. 2020). An emerging policy paradigm focuses on societal objectives, namely, how new technologies can help to address societal challenges such as climate change, growing inequality, demographic change or resource scarcity (Diercks, Larsen, and Steward 2019). It takes a broader approach to the innovation process including not only industry, academia and government but also civil society and considering not only supply-side of technology but also demand-side of its uses (Diercks, Larsen, and Steward 2019). Of course, many of these ideas are not completely new and some of them can be traced back to the long-standing work on social function of science (Bernal 1939), the Moon and the Ghetto debate on differences in technological progress in different domains (Nelson 1977, 2011), and efforts to prioritize social justice issues in research and development policy-making agendas (Woodhouse and Sarewitz 2007). While societal challenges such as climate change are global in their nature, it is also important to contextualize them and recognize their plurality (Wanzenbock et al. 2020). This new transformative innovation policy paradigm, which brings these ideas together in a systematic framework, can be relevant for AI policy.

Beyond global competition: international research collaboration and science diplomacy

If many countries and international organizations have developed their national AI strategies, then how do they interact with each other? Policy documents and media promote seemingly contradictory discourses of global competition vs. global cooperation on AI (Ulnicane et al. forthcoming). The global competition discourse presents AI development as ‘a new space race’ where

countries compete to dominate the AI development. This discourse focuses on which countries are making major investments and advances in AI. An example of this global competitiveness narrative can be seen in this quote from the UK document stating that

A challenging race to make most of the opportunities posed by AI has begun. China, the US, Russia, Canada, Japan, and many more countries have passed ambitious strategies in which they have put AI as a priority in their political agendas. (BIC/APPGAI 2017b, 5)

A number of countries and organizations have expressed their ambitions to be global leaders and defined their approaches to global competition, for example, the European Union has announced that it wants to lead based on its values (Ulnicane 2021). This global competitiveness discourse demonstrates the priority countries and organizations assign to AI but it can also have negative effects such as hampering global collaboration. Focus on global competitiveness among countries has been called ‘a dangerous obsession’ (Krugman 1994) because it suggests that relationships among countries are ‘a zero-sum game’ where one country wins and other loses rather than ‘a positive-sum game’ where the overall size of the pie increases and everyone gains.

At the same time, there are many calls for global cooperation and coordination on AI development and internationally recognized ethical and legal frameworks to address common challenges (e.g. EGE 2018; Rathenau Institute 2017). This quote from the EU High-Level Expert Group on AI summarizes some of the main arguments for global cooperation:

Just as the use of AI systems does not stop at national borders, neither does their impact. Global solutions are therefore required for the global opportunities and challenges that AI systems bring forth. We therefore encourage all stakeholders to work towards a global framework for Trustworthy AI, building international consensus while promoting and upholding our fundamental rights-based approach. (European Commission 2019, 5)

International cooperation on AI is already taking place in the European Union, OECD, G20 and other organizations (e.g. European Commission 2020). An important part of AI policy development around the world is that countries learn from each other’s AI initiatives and strategies (e.g. BIC/APPGAI 2017b) and such cross-country learning is facilitated by international forums such as the World Economic Forum. Suggestions have been made to launch international cooperation initiatives for AI similar to success stories in other fields such as the Intergovernmental Panel on Climate Change (European Commission 2018b) and the European Organization for Nuclear Research CERN for AI experts to discuss and develop technology (European Commission 2017).

The development of international cooperation in AI can benefit from lessons learned on global cooperation among researchers and science diplomacy among policy-makers in other fields. Extensive social science studies of

international scientific collaboration demonstrate the increase in scientists collaborating across national borders due to various reasons such as bringing together diverse types of expertise, seeking reputation, pooling resources for large-scale infrastructure and addressing cross-border problems (e.g. Wagner, Whetsell, and Mukherjee 2019). While such collaborations can bring many benefits, they also encounter difficulties such as high transaction costs due to diverse cultures and long geographical distances (Wagner, Whetsell, and Mukherjee 2019). To facilitate international collaborations among scientists, policy-makers have launched science diplomacy activities at the intersection of science policy and foreign affairs to support joint efforts to address global challenges (e.g. Flink and Rüffin 2019). When designing international cooperation initiatives for AI, which have to address not only complex scientific and technological issues but also often sensitive topics of diverse political and economic systems, regulatory environments and cultural traditions, there are many opportunities to learn what works in other fields and how. Does AI really need a large-scale, single-site physical research infrastructure such as CERN, or would a networked and distributed collaboration be more appropriate in this area?

Beyond biomedical ethics: ethics of computing

The dominant approach to ethics as applied to AI is based on biomedical ethics (see e.g. Mittelstadt 2019). Biomedical ethics arose following the second World War and the Nazi atrocities committed in the concentration camps. The principles were formed during the Nuremberg trials of the war criminals (Freyhofer 2004). These were developed in the World Medical Association's (2008) Helsinki Declaration and formalized through the Belmont report (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979). A cornerstone of biomedical ethics is its reliance on mid-level ethical principles (beneficence, non-maleficence, justice and autonomy) (Beauchamp and Childress 2009). This approach of basing applied ethics on mid-level principles is dominant in AI ethics, as Anna Jobin and colleagues' (2019) review shows.

While the adoption of biomedical ethics offers advantages, notably the immediate recognition of approach and the possibility of utilizing existing structures of biomedical ethics (institutional review boards, research ethics committees, established review processes), it is important to point out that the 'principilism' (Clouser and Gert 1990) of biomedical ethics has always been subject to controversy. Even from within the biomedical field it has been regarded as overly rigid (Klitzman 2015) and there have been questions of its consistency and applicability beyond the biomedical field (Schrag 2010; Stark 2011). A key concern with high relevance to AI is that biomedical ethics is based on the assumption that the underlying research is fundamentally ethically desirable

(understanding disease, finding cures) and the main point of ethics is the protection of patients. While this assumption holds for much of biomedical research, it is arguably much weaker elsewhere, including in AI, where it is not *a priori* obvious that a new technology or innovation *per se* is desirable.

In addition, the focus on biomedical ethics renders invisible a large body of work undertaken on ethical questions of computing. This work started in the early days of digital computing (Wiener 1954) and has become a more formal sub-discipline with dedicated journals and conferences since the 1980s (Moor 1985). There are bodies of work around concepts such as computer ethics (Gotterbarn 1995; Johnson 2001), information ethics (Floridi 1999), digital ethics and others, which are dedicated to ethical aspects of information and communication technologies. Importantly, a dedicated sub-field of Inter-cultural Digital Ethics brings in diverse cultural and social perspectives to examine ethical issues of digital technologies (Aggarwal 2020). AI, traditionally classified as one field of computer science, has been subject of these studies well before the current AI publicity. However, by relying on biomedical ethics, much of this work has been rendered less visible than it arguably deserves to be, thus limiting its ability to contribute to making AI beneficial.

Many issues at the core of current AI debates such as privacy, autonomy, agency, trust and inclusion have been extensively addressed in the field of computing ethics (Stahl, Timmermans, and Mittelstadt 2016), which have demonstrated the breadth of influence that technology can have on all aspects of life affecting power and politics, economics, education, the environment and other areas. As there is a lot of continuity and path-dependence from computing to AI technologically as well as in terms of its wide-ranging impacts, AI can benefit by building on the work done and knowledge accumulated on ethical and societal issues of computing.

Beyond ethical principles: Responsible Innovation

In the past decade the Responsible Innovation approach has been developed as a way to go beyond ethical principles (Owen and Pansera 2019). It has been applied in particular in Europe over a range of emerging technosciences such as nanotechnology, geoengineering and synthetic biology. While there are many definitions of Responsible Innovation, it is broadly understood that

responsible forms of innovation should be aligned to social needs, be responsive to changes in ethical, social and environmental impacts as a research programme develops, and include the public as well as traditionally defined stakeholders in two-way consultation. (De Saille 2015, 153)

The alignment of innovation to societal needs includes processes and practices of anticipation of potential future developments and impacts of science and innovation, inclusion of diverse stakeholders, reflexivity and responsiveness

(Stilgoe, Owen, and Macnaghten 2013). While the specific term of Responsible Innovation has emerged only about a decade ago, it builds on long-established practices such as technology assessment and public engagement (e.g. Jasanoff 2016). What is novel about the RRI approach is that it aims to bring such practices into a more systematic framework along the four interconnected dimensions of anticipation, inclusion, reflexivity and responsiveness (Stilgoe, Owen, and Macnaghten 2013). Each of these four dimensions offers of a number of techniques to address and implement societal responses into research and innovation. Anticipation dimension includes techniques of foresight, technology and risk assessment, inclusion dimension encompasses approaches such as citizen conferences and focus groups, reflexivity dimension includes multidisciplinary training and collaboration as well as embedded social scientists and ethicists in laboratories, while responsiveness dimension draws on the insights from anticipation, inclusion and reflexivity exercises to design appropriate measures from regulation and funding programmes to moratoriums if need be. This can be a powerful approach to address societal needs by systematically integrating anticipation, inclusion, reflexivity and responsiveness activities and involving a broad range of stakeholders in them.

Some AI policy documents mention Responsible AI without explaining it and making explicit links to the Responsible Innovation approach. Several documents mention Responsible Innovation in passing with one exception (IEEE 2017) of engaging with Responsible Innovation work and its relevance for AI. More in-depth learning from a decade of Responsible Innovation advances and limitations in aligning emerging technosciences with societal needs could contribute to developing systematic frameworks for building socially beneficial AI. So far in other fields Responsible Innovation approach has been mostly applied at the level of research projects and laboratories where, for example, social scientists and ethicists collaborate with technology developers in knowledge co-production (see e.g. Aicardi et al. 2020) but it also has a potential to consider the role of politics and power to further democratize technology development and use (Van Oudheusden 2014). In that way, project and laboratory level Responsible Innovation practices can benefit from being part of a broader governance processes described in previous lessons such as diverse roles of state at national level, new policy paradigms of societal challenges and international cooperation arrangements.

To summarize, the six above-discussed lessons from the literature on governance, policy and ethics of emerging technologies are complementary and reinforcing each other. Insights from the governance literature focus on involving diverse stakeholders in decision-making, which can be facilitated by the state and organized along the dimensions of Responsible Innovation which focus on anticipation, inclusion, reflexivity and responsiveness. These can be geared towards developing and using AI in a way that addresses societal challenges locally and globally. They offer plenty of opportunities to develop many

concrete recommendations for governance of socially beneficial AI. However, concrete recommendations cannot be produced in a universal ‘one size fits all’ manner. Rather they should be developed in a reflexive, collaborative and inclusive way considering specific contexts, cultures and traditions.

Conclusions: some lessons for a good governance for AI

Recent advances in AI have triggered intensive public debates about potential impact of this technology. Against this background, AI policy documents launched by national governments, international organizations and various stakeholders focus on wide-ranging and long-lasting effects on jobs, politics, economy and society. These effects include positive expectations as well as major concerns about AI effects on individual and societal level. While some of these concerns, such as effects of automation, might seem novel, others related to risks and inequalities are well-known. Typical solutions for dealing with these concerns suggested in AI policy documents focus on ethics and regulation. The role of both of these suggested solutions remains contested.

To look for systematic ways to address concerns associated with AI and facilitate its development and use in socially beneficial ways, in this article we reviewed some recent lessons from literature on governance, policy and ethics of emerging technologies. Such lessons can be relevant for AI which share typical features of emerging technologies such as fast growth, radical novelty, prominent impact and uncertainty. We derive six complementary and reinforcing lessons. First, rather than focus on top-down government decisions, consider governance arrangements that bring together government and diverse non-state groups from civil society and private sector in a balanced and transparent way. Second, rather than assume limited role of the state in market correction, think about diverse roles of the state in mitigating risks, enabling participation of diverse groups and mediating various needs and interests. Third, go beyond traditional goals of technology to support economic growth and productivity, and focus on how technology can address societal challenges. Fourth, rather than being a global race where one country wins and others loose, technology development can be based on international research collaboration supported by science diplomacy efforts. Fifth, rather than develop AI ethics based on principles of medical ethics, learn from a related field of computing ethics that has accumulated extensive knowledge on issues such as privacy, autonomy, agency, trust and inclusion. Sixth, one way to go beyond ethical principles is to learn from the Responsible Innovation approach on how to systematically address societal needs in the development and use of emerging technologies.

Governance of emerging technologies is a highly complex endeavour and there are no magic solutions. However, to address concerns associated with AI and shape its development in socially beneficial ways, recent lessons from

other emerging technologies can offer ideas, concepts and approaches that broaden a range of available options and allow to imagine various ways of achieving desirable objectives. Further work should focus on examining governance needs and arrangements of specific AI applications in diverse contexts at different but interconnected levels from technology development projects and laboratories to national policies and international cooperation arrangements.

Acknowledgements

This paper has benefited from the comments and suggestions on an earlier version presented at the Science in Public 2018 conference in Cardiff, Wales. The authors are grateful to Tonii Leach, Dinesh Mothi and Winter-Gladys Wanjiku who contributed to the document analysis as well as to Juliana Nnadi and Iffat Islam who participated in early discussions on the framing of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 720270 (HBP SGA1), No. 785907 (HBP SGA2) and No. 945539 (HBP SGA3).

Notes on contributors

Dr Inga Ulnicane has extensive international and interdisciplinary experience of research, teaching and engagement in the field of science, technology and innovation governance. Her scientific publications and commissioned reports focus on topics such as research and innovation policies, international research collaboration, Grand societal challenges, and dual use. She has worked at University of Vienna (Austria), University of Twente (Netherlands), University of Latvia and Latvian Academy of Sciences, and has been visiting scientist at University of Manchester (UK) and Georgia Institute of Technology (US). Currently she is at De Montfort University (UK).

Damian Okaibedi Eke (PhD) is a Research Fellow in the EU Human Brain Project in the Centre for Computing and Social Responsibility at De Montfort University, UK. Damian has Computer Ethics background and his current research includes work on responsible data governance of biomedical data, Data Ethics, Ethics of Emerging technologies including AI and ICT4D.


Dr William Knight is a research fellow at the Centre for Computing and Social Responsibility at De Montfort University, Leicester, UK. His research interests include health research, hacking and online activism, research management and ethics compliance, data governance and data protection. Dr Knight is the ethics compliance manager for EU funded Future Emerging Technology flagship: Human Brain Project.


George Ogoh is a Research Fellow at the Centre for Computing and Social Responsibility, De Montfort University. His current role is in the Human Brain Project (HBP) – an EU funded Future Emerging Technology flagship. His research interests cover a wide range of topics in emerging technology ethics including responsible innovation, data governance and data protection.

Bernd Carsten Stahl is Professor of Critical Research in Technology and Director of the Centre for Computing and Social Responsibility at De Montfort University, Leicester, UK. His interests cover philosophical issues arising from the intersections of business, technology, and information. This includes ethical questions of current and emerging information and communication technologies, critical approaches to information systems and issues related to responsible research and innovation.


ORCID

Inga Ulnicane  <http://orcid.org/0000-0003-2051-1265>

Damian Okaibedi Eke  <http://orcid.org/0000-0002-6210-1283>

William Knight  <http://orcid.org/0000-0001-9818-6277>

George Ogoh  <http://orcid.org/0000-0002-5287-408X>

Bernd Carsten Stahl  <http://orcid.org/0000-0002-4058-4456>

References

- The 2015 panel. 2016. “Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence.” Report of the 2015 Study Panel.
- Accenture. 2017. “Embracing Artificial Intelligence.” Enabling Strong and Inclusive AI Driven Growth.
- Adams, Rachel. 2021. “Decolonising Artificial Intelligence: A Theoretical Critique.” *Interdisciplinary Science Reviews* 46 (1).
- Aggarwal, Nikita. 2020. “Introduction to the Special Issue on Intercultural Digital Ethics.” *Philosophy & Technology* 33 (4): 547–550. doi:10.1007/s13347-020-00428-1.
- Aicardi, Christine, Simisola Akintoye, B. Tyr Fothergill, Manuel Guerrero, Gudrun Klinker, William Knight, Lars Klüver, et al. 2020. “Ethical and Social Aspects of Neurorobotics.” *Science and Engineering Ethics* 26 (5): 2533–2546. doi:10.1007/s11948-020-00248-8.
- Beauchamp, Tom L., and James F. Childress. 2009. *Principles of Biomedical Ethics*. 6th ed. New York: OUP USA.
- Bernal, John D. (1939) 1967. *The Social Function of Science*. Cambridge: The MIT Press.
- BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence). 2017a. Governance, Social and Organisational Perspective for AI. 11 September 2017.
- BIC/APPGAI (Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence). 2017b. International Perspective and Exemplars. 30 October 2017.
- Blackwell, Alan. 2021. “Ethnographic Artificial Intelligence.” *Interdisciplinary Science Reviews* 46 (1).
- Borras, Susana, and Jakob Edler, eds. 2014. *The Governance of Socio-Technical Systems: Explaining Change*. Cheltenham: Edward Elgar.
- Borras, Susana, and Jakob Edler. 2020. “The Roles of the State in the Governance of Socio-Technical Systems’ Transformation.” *Research Policy* 49 (5): 103971.

- Bowser, Anne, Michael Sloan, Pietro Michelucci, and Eleonore Pauwels. 2017. *Artificial Intelligence: A Policy-Oriented Introduction*. Washington, DC: Wilson Briefs. Wilson Center.
- Campolo, Alex, Medelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. *AI Now 2017 Report*. AI Now Institute. New York: New York University.
- Clouser, K. Danner, and Bernard, Gert. 1990. "A Critique of Principlism." *Journal of Medicine and Philosophy* 15: 219–236. doi:10.1093/jmp/15.2.219.
- Coeckelbergh, Mark, and Thomas Metzinger. 2020. "Europe Needs More Guts When it Comes to AI Ethics." *Tagesspiegel*, April 14.
- Crawford, Kate, and Meredith Whittaker. 2016. *The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. New York: AI Now Institute.
- Datta Burton, Saheli, Tara Mahfoud, Christine Aicardi, and Nikolas Rose. 2021. "Clinical translation of computational brain models: understanding the salience of trust in clinician-researcher relationships." *Interdisciplinary Science Reviews* 46 (1).
- De Saille, Stevienna. 2015. "Innovating Innovation Policy: The Emergence of 'Responsible Research and Innovation'." *Journal of Responsible Innovation* 2 (2): 152–168. doi:10.1080/23299460.2015.1045280.
- De Saille, Stevienna, Fabien Medvecky, Michiel van Oudheusden, Kevin Albertson, Effie Amanatidou, Timothy Birabi, and Mario Pantera. 2020. *Responsibility Beyond Growth. A Case for Responsible Stagnation*. Bristol: Bristol University Press.
- Diercks, Gijs, Henrik Larsen, and Fred Steward. 2019. "Transformative Innovation Policy: Addressing Variety in an Emerging Policy Paradigm." *Research Policy* 48 (4): 880–894.
- EESC (European Economic and Social Committee). 2017. "Artificial Intelligence - the Consequences of Artificial Intelligence on the (Digital) Single Market, Production, Consumption, Employment and Society." *Opinion*.
- EGE (European Group on Ethics in Science and New Technologies). 2018. "Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems."
- European Commission. 2017. *AI Policy Seminar: Towards and EU Strategic Plan for AI*. Brussels: Digital Transformation Monitor.
- European Commission. 2018a. *Artificial Intelligence: A European Perspective*. Luxembourg: Publications Office of the European Union.
- European Commission. 2018b. "Coordinated Plan on Artificial Intelligence." Communication COM(2018) 795 final. Brussels 7.12.2018.
- European Commission. 2018c. "Artificial Intelligence for Europe. Communication." COM (2018) 237 final. Brussels 25.4.2018.
- European Commission. 2019. "Ethics Guidelines for Trustworthy AI. Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission." Brussels. 8.4.2019.
- European Commission. 2020. "On Artificial Intelligence – A European Approach to Excellence and Trust." White Paper. COM(2020) 65 final. Brussels 19.2.2020.
- European Parliament. 2016. "European Civil Law Rules in Robotics." Study for the JURI Committee.
- Executive Office of the President. 2016a. *The National Artificial Intelligence Research and Development Strategic Plan*. Washington, DC: National Science and Technology Council.
- Flink, Tim, and Nicolas Rüffin. 2019. "The Current State of the Art of Science Diplomacy." In *Handbook on Science and Public Policy*, edited by Dagmar Simon, Stefan Kuhlmann, Julia Stamm, and Weert Canzler, 104–121. Cheltenham: Edward Elgar.
- Floridi, Luciano. 1999. "Information Ethics: On the Philosophical Foundation of Computer Ethics." *Ethics and Information Technology* 1: 33–52.

- Freyhofer, Horst H. 2004. *The Nuremberg Medical Trial: The Holocaust and the Origin of the Nuremberg Medical Code*, 2nd Revised ed. New York: Peter Lang.
- Garvey, Colin Shunryu. 2021. "Unsavoury Medicine for Technoscientific Civilization: Introduction to AI & its Discontents." *Interdisciplinary Science Reviews* 46 (1).
- Gotterbarn, Donald. 1995. "Computer Ethics – Responsibility Regained." In *Computers, Ethics and Social Values*, edited by Deborah G. Johnson and Helen Nissenbaum, 18–24. Upper Saddle River: Prentice Hall.
- Hagendorff, Thilo. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30: 99–120. doi:10.1007/s11023-020-09517-8.
- Holzmeyer Cheryl. 2021. "Beyond "AI for Social Good" (AI4SG): Social Transformations - Not Tech-Fixes - for Health Equity." *Interdisciplinary Science Reviews* 46 (1).
- IEEE. 2017. "Ethically Aligned Design. A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems." Version 2 – for Public Discussion.
- ITU (International Telecommunication Union). 2017. *AI for Good Global Summit Report 2017*, Geneva, 7–9 June 2017.
- Jasanoff, Sheila. 2016. *The Ethics of Invention: Technology and the Human Future*. New York: W.W. Norton.
- Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1: 389–399. doi:10.1038/s42256-019-0088-2.
- Johnson, Deborah G. 2001. *Computer Ethics*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.
- Juma, Calestous. 2016. *Innovation and Its Enemies. Why People Resist New Technologies*. Oxford: Oxford University Press.
- Klitzman, Robert. 2015. *The Ethics Police?: The Struggle to Make Human Research Safe*. 1st ed. Oxford; New York: OUP USA.
- Krugman, Paul. 1994. "Competitiveness: A Dangerous Obsession." *Foreign Affairs* 73 (2): 28–44.
- Kuhlmann, Stefan, Peter Stegmaier, and Kornelia Konrad. 2019. "The Tentative Governance of Emerging Science and Technology – a Conceptual Introduction." *Research Policy* 48 (5): 1091–1097. doi:10.1016/j.respol.2019.01.006.
- Loeb, Zachary. 2021. "The Lamp and the Lighthouse: Joseph Weizenbaum, contextualizing the critic." *Interdisciplinary Science Reviews* 46 (1).
- Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence we Can Trust*. New York: Pantheon Books.
- Metzinger, Thomas. 2019. "Ethics Washing Made in Europe." *Tagesspiegel*, April 8.
- Mittelstadt, Brent. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1 (11): 501–507. doi:10.1038/s42256-019-0114-4.
- Moor, James H. 1985. "What is Computer Ethics." *Metaphilosophy* 16: 266–275.
- The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report - Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, DC: Department of Health, Education, and Welfare.
- Nelson, R. Richard. 1977. *The Moon and the Ghetto. An Essay on Public Policy Analysis*. New York: W.W. Norton & Company.
- Nelson, R. Richard. 2011. "The Moon and The Ghetto Revisited." *Science and Public Policy* 38 (9): 681–690.
- Ochigame, Rodrigo. 2019. "How Big Tech Manipulates Academia to Avoid Regulation." *The Intercept*, December 20. <https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/>.
- Olson, Mancur. 1974. *The Logic of Collective Action: Public Goods and the Theory of Groups*. 2nd ed. Cambridge, MA: Harvard University Press.

- Owen, Richard., and Mario. Pansera. 2019. "Responsible Innovation and Responsible Research and Innovation." In *Handbook on Science and Public Policy*, edited by Dagmar. Simon, Stefan. Kuhlmann, Julia. Stamm, and Weert. Canzler, 26–48. Cheltenham: Edward Elgar.
- Pierre, Jon, and B. Guy Peters. 2000. *Governance, Politics and the State*. Houndmills: Macmillan.
- Rathenau Institute. 2017. "Human Rights in the Robot Age. Challenges Arising from the Use of Robotics, Artificial Intelligence, and Virtual and Augmented Reality." Report for the Parliamentary Assembly of the Council of Europe.
- Rotolo, Daniele, Diana Hicks, and Ben R. Martin. 2015. "What is an Emerging Technology?" *Research Policy* 44 (10): 1827–1843. doi:10.1016/j.respol.2015.06.006.
- Schrag, Zachary. M. 2010. *Ethical Imperialism: Institutional Review Boards and the Social Sciences, 1965–2009*. 1st ed. Baltimore: Johns Hopkins University Press.
- Schwab, Klaus. 2017. *The Fourth Industrial Revolution*. London: Penguin.
- Stahl, Bernd Carsten, Job Timmermans, and Brent Daniel Mittelstadt. 2016. "The Ethics of Computing: A Survey of the Computing-Oriented Literature." *ACM Computing Surveys* 48 (4): 55:1–55:38. doi:10.1145/2871196.
- Stark, Laura. 2011. *Behind Closed Doors: IRBs and the Making of Ethical Research*. 1st ed. Chicago: University of Chicago Press.
- Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580. doi:10.1016/j.respol.2013.05.008.
- Thierer, Adam, Andrea Castillo O'Sullivan, and Raymond Russell. 2017. *Artificial Intelligence and Public Policy. Report*. Arlington: Mercatus Center, George Mason University.
- Ulnicane, Inga, William Knight, Tonii Leach, Bernd Carsten Stahl, and Winter-Gladys Wanjiku. 2020. Framing Governance for a Contested Emerging Technology: Insights from AI Policy. *Policy and Society*. doi:10.1080/14494035.2020.1855800.
- Ulnicane, Inga. 2021. "Artificial Intelligence in the European Union: Policy, Ethics and Regulation." In *Routledge Handbook of European Integrations*, edited by Thomas Hoerber, Ignazio Cabras, and Gabriel Weber. Routledge.
- Ulnicane, Inga, William Knight, Tonii Leach, Bernd Carsten Stahl, and Winter-Gladys Wanjiku. Forthcoming. "Governance of Artificial Intelligence: Emerging International Trends and Policy Frames." In *Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello. Boca Raton: CRC Press.
- UNI Global Union. 2017. "Top 10 Principles for Ethical Artificial Intelligence." The Future World of Work.
- Van Oudheusden, Michiel. 2014. "Where Are the Politics in Responsible Innovation? European Governance, Technology Assessments, and Beyond." *Journal of Responsible Innovation* 1 (1): 67–86. doi:10.1080/23299460.2014.882097.
- Vesnic-Alujevic, Lucia, Susana Nascimento, and Alexandre Polvora. 2020. "Societal and Ethical Impacts of Artificial Intelligence: Critical Notes on European Policy Frameworks." *Telecommunications Policy* 44 (6): 101961.
- Villani, Cedric. 2018. *For a Meaningful Artificial Intelligence. Towards a French and European Strategy*. Paris.
- Vinnova. 2018. *Artificial Intelligence in Swedish Business and Society*. Stockholm: Vinnova.
- Wagner, Caroline, Travis Whetsell, and Satyam Mukherjee. 2019. "International Research Collaboration: Novelty, Conventionality, and Atypicality in Knowledge Recombination." *Research Policy* 48: 1260–1270.

- Wanzenböck, Iris, Joeri H. Wesseling, Koen Frenken, Marko P. Hekkert, and K. Matthias Weber. 2020. "A Framework for Mission-Oriented Innovation Policy: Alternative Pathways Through the Problem-Solution Space." *Science and Public Policy*. doi:10.1093/scipol/scaa027.
- Wiener, Norbert. 1954. *The Human Use of Human Beings*. New York: Doubleday.
- Woodhouse, Edward J., and Daniel Sarewitz. 2007. "Science Policies for Reducing Societal Inequities." *Science and Public Policy* 34 (2): 139–150.
- World Medical Association. 2008. "Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects."

Appendix

AI policy documents analysed (in alphabetical order)

1. Accenture (2017) Embracing artificial intelligence. Enabling strong and inclusive AI driven growth.
2. Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence (2017a) APPG AI Findings 2017.
3. Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence (2017b) Governance, Social and Organisational Perspective for AI. 11 September 2017.
4. Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence (2017c) Inequality, Education, Skills, and Jobs. 16 October 2017.
5. Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence (2017d) International Perspective and Exemplars. 30 October 2017.
6. Big Innovation Centre/All-Party Parliamentary Group on Artificial Intelligence (2017e) What is AI? A theme report based on the 1st meeting of the All-Party Parliamentary Group on Artificial Intelligence. 20 March 2017.
7. Bowser, A., M. Sloan, P. Michelucci and E. Pauwels (2017) Artificial Intelligence: A Policy-Oriented Introduction. Wilson Briefs. Wilson Center.
8. Campolo, A, M.Sanfilippo, M.Whittaker and K.Crawford (2017) AI Now 2017 Report. AI Now Institute, New York University.
9. CNIL (2017) Algorithms and artificial intelligence: CNIL's report on the ethical issues.
10. Crawford, K. and M.Whittaker (2016) The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term. AI Now Institute.
11. EDPS (2016) Artificial Intelligence, Robotics, Privacy and Data Protection. Room document for the 38th International Conference of Data Protection and Privacy Commissioners.
12. European Commission (2017) AI Policy Seminar: Towards and EU strategic plan for AI. Digital Transformation Monitor.
13. European Commission (2018a) Artificial Intelligence: A European Perspective.
14. European Commission (2018b) Artificial Intelligence for Europe. Communication.
15. European Commission (2018c) Coordinated Plan on Artificial Intelligence. Communication.
16. European Economic and Social Committee (2017) Artificial Intelligence - The consequences of Artificial intelligence on the (digital) single market, production, consumption, employment and society. Opinion.
17. European Group on Ethics in Science and New Technologies (2018) Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems.

18. European Parliament (2016) European Civil Law Rules in Robotics. Study for the JURI Committee.
19. European Parliament (2017) Report with recommendations to the Commission on Civil Law Rules on Robotics.
20. European Parliament (2018) Understanding Artificial Intelligence. Briefing EPRS.
21. Executive Office of the President (2016a) Artificial Intelligence, Automation, and Economy, Report.
22. Executive Office of the President (2016b) Preparing for the future of artificial intelligence. National Science and Technology Council Committee on Technology.
23. Executive Office of the President (2016c) The National Artificial Intelligence research and development Strategic Plan. National Science and Technology Council. Networking and Information Technology Research and Development Subcommittee.
24. Future of Humanity Institute et al (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation.
25. Government Office for Science (2016) Artificial Intelligence: opportunities and implications for the future of decision making.
26. Hall, W. and J. Pesenti (2017) Growing the Artificial Intelligence Industry in the UK.
27. HM Government (2018) Artificial Intelligence Sector Deal. 26 April 2018.
28. House of Commons Science and Technology Committee (2016) Robotics and artificial intelligence. Fifth report of session 2016-17.
29. House of Lords (2018) AI in the UK: ready, willing and able?
30. IEEE (2017) Ethically aligned design. A vision for prioritizing human well-being with autonomous and intelligent systems. Version 2 – for public discussion.
31. IEEE European Public Policy Initiative (2017) Artificial Intelligence: Calling on Policy--Makers to Take a Leading Role in Setting a Long-Term AI Strategy. Position Statement.
32. IEEE-USA (2017) Artificial Intelligence Research, Development & Regulation. Position Statement.
33. Information Commissioner's Office (2017) Big data, artificial intelligence, machine learning and data protection. Data Protection Act and General Data Protection Regulation.
34. International Telecommunication Union (2017) AI for Good Global Summit Report 2017, Geneva, 7-9 June 2017.
35. IPPR (2017) Managing automation: Employment, inequality and ethics in the digital age. Discussion Paper.
36. Ministry of Economic Affairs and Employment (2017) Finland's Age of Artificial Intelligence.
37. Ponce Del Castillo, A. (2017) A Law on Robotics and Artificial Intelligence in the EU? Foresight Brief. European Trade Union Institute ETUI.
38. Rathenau Institute (2017) Human Rights in the Robot Age. Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. Report for the Parliamentary Assembly of the Council of Europe.
39. SGPAC (2017) Governance, Risk & Control: Artificial Intelligence. Effective Deployment, Management and Oversight of Artificial Intelligence (AI). Version 1.0. 22 March 2017. SGPAC Consulting & Advisory.
40. Tata Leading the Way with Artificial Intelligence: The Next Big Opportunity for Europe. TCS Global Trend Study – Europe. Tata Consultancy Services.
41. The 2015 panel (2016) Artificial Intelligence and life in 2030. One hundred year study on artificial intelligence. Report of the 2015 study panel.
42. The Federal Government (2018) Artificial Intelligence Strategy. November 2018
43. The Royal Society (2017) Machine learning: the power and promise of computers that learn by example.

44. Thierer, A., A. Castillo O'Sullivan, and R. Russell (2017) Artificial Intelligence and Public Policy. Report. Mercatus Center, George Mason University.
45. UNI Global Union (2017) Top 10 Principles for ethical artificial intelligence. The future world of work.
46. Villani, C. (2018) For a meaningful artificial intelligence. Towards a French and European Strategy.
47. Vinnova (2018) Artificial Intelligence in Swedish business and society.
48. Whittaker, M., K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. Myers West, R. Richardson, J. Schultz, O. Schwartz (2018) AI Now Report 2018.
49. World Economic Forum (2018) Artificial Intelligence for the Common Good. Sustainable, Inclusive and Trustworthy. White Paper for attendees of the WEF 2018 Annual Meeting.



Beyond 'AI for Social Good' (AI4SG): social transformations—not tech-fixes—for health equity

Cheryl Holzmeyer

Institute for Social Transformation, University of California, Santa Cruz, CA, USA

ABSTRACT

This paper reflects on proliferating AI for Social Good (AI4SG) initiatives, with an eye to public health and health equity. It notes that many AI4SG initiatives are shaped by the same corporate entities that incubate AI technologies, beyond democratic control, and stand to profit monetarily from their deployment. Such initiatives often pre-frame systemic social and environmental problems in tech-centric ways, while suggesting that addressing such problems hinges on more or better data. They thereby perpetuate incomplete, distorted models of social change that claim to be 'data-driven'. In the process, AI4SG initiatives may obscure or 'ethics wash' all the other uses of big data analytics and AI that more routinely serve private interests and exacerbate social inequalities. As a case in point, it discusses the prominence of health-related applications in AI and big data fields, alongside the politics of more 'upstream' versus 'downstream' health interventions.

KEYWORDS

Artificial Intelligence (AI); AI for Social Good (AI4SG); AI Ethics; AI Governance; Precision Medicine; Social Justice; Public Health; Health Equity

Introduction

As the Covid-19 pandemic alters everyday landscapes of possibility around the globe – in so many disparate, fractured ways – another kind of emerging landscape is being articulated under the umbrella of 'Artificial Intelligence (AI) for Social Good' (Tomašev, Cornebise, and Hutter 2020). Championed by stakeholders from Big Tech companies to corporate management consultancies to the United Nations (UN), AI for Social Good (AI4SG) projects (or sometimes simply, AI for Good) emphasize computation and the deployment of big data analytics, including machine learning, to address a wide range of social and environmental issues. While AI4SG projects are often just another way to frame the AI activities of conventional profit-oriented business entities, the formal field of AI4SG is gathering momentum, basking in the glow of an apparent 'AI summer'. A recent survey of the field found over 1000 published papers on AI4SG topics (Shi, Wang, and Fang 2020, 1), growing from 18 papers in

CONTACT Cheryl Holzmeyer  cholzme@ucsc.edu  Institute for Social Transformation, University of California, Santa Cruz, CA, USA

© 2020 Institute of Materials, Minerals and Mining Published by Taylor & Francis on behalf of the Institute

2008 to 246 papers in 2019 (Shi, Wang, and Fang 2020, 5). AI4SG projects might identify endangered species in digital image streams to aid conservation efforts; they might analyze satellite data to monitor manifestations of climate change, such as sea-level rise or desertification; or they might support health diagnostics, for instance, by detecting skin cancers based on mobile phone photos (Chui et al. 2018). They might help with Covid-19 contact tracing or symptom tracking. Such projects and their claims to ‘do good’ are part of the broader landscape of efforts to define responsible innovation and ethical AI (Ulnicane et al., this issue), with the vast majority of formal, non-binding ethics guidelines articulated in a handful of the world’s wealthiest countries (Jobin, Ienca, and Vayena 2019, 391). AI4SG is also part of the longer lineage of ‘Technology for Good’ initiatives, including the field of Information and Communication Technologies for Development.

This essay reflects on proliferating AI4SG initiatives, with an eye to public health and health equity in a US context, engaging questions that emerged in part from the author’s experiences with projects that attempted to leverage digital technological infrastructures, including big data analytics, on behalf of public health (Holzmeyer 2018a, 2018b, 2018c). It draws as well on a broader literature review and consideration of existing political, economic and social structures shaping ‘innovation’ in relation to health equity, meaning a society in which ‘*everyone* has the opportunity to attain their highest level of health’ (American Public Health Association; emphasis added). It highlights ways in which existing metrics and goals of innovation, within and beyond AI, such as patents and GDP growth, are deeply inadequate (or even regressive) as indicators of and guides to a flourishing, inclusive, sustainable, democratic society (Mazzucato 2018). It asserts that, indeed, Another World is Possible (contra Margaret Thatcher’s ‘TINA’ claim that There Is No Alternative), beyond the confines of the neoliberal market fundamentalism that for decades has been integral to policies shredding social safety nets and the underpinnings of well-being for so many people and ecosystems around the world (Beckfield 2018; Coburn 2010; Wilkinson and Pickett 2019), including in a supposedly innovative and materially wealthy society such as the USA. These social and environmental justice issues should be integral to science and technology policy-making going forward, with health equity as a touchstone for innovation and STEM (Science, Technology, Engineering, and Mathematics) education (Holzmeyer 2017), in order to realize the vision articulated in the World Health Organization (WHO) Constitution of ‘the highest attainable standard of health as a fundamental right of every human being’, with health described in the preamble as ‘a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity’ (WHO 2017).

Moreover, whether or not they are explicitly engaged with conversations around emerging technology governance, the social movements that are currently demonstrating in the USA and beyond for racial, economic, climate

and other forms of social and environmental justice are true thought leaders in imagining the new social compacts that should be the baselines and foundations of democratic AI governance, including of any empirically reasonable claims of ‘AI for Social Good’ (and various leading organizations, such as Data for Black Lives [<https://d4bl.org/>], are engaged in both social movement activism and technology governance conversations; c.f. Milner 2018, 2020). This social movement thought leadership is echoed in statements by progressive academic scholars in fields from economics (Sachs et al. 2020) to public health (Bassett 2015; Krieger 2020), if not as yet in any thoroughgoing, institutionalized way in Big Tech’s AI public engagement initiatives (with ‘solidarity’ being one of the underrepresented values in AI ethics guidelines to date; c.f. Jobin, Ienca, and Vayena 2019, 396).

This paper proceeds with an examination of the ambiguous meanings of AI and AI4SG – and the politics, harmful externalities and alternative social worlds that may be obscured and foreclosed by the ‘social good’ presumptions of AI4SG. As a case in point, it discusses the prominence of health-related applications in AI and big data fields in the USA, including in avowed AI4SG projects. It examines the politics of different types of health interventions, particularly the tension between growing investments in ever more individualized treatments in downstream clinical settings, on the one hand, and underinvestment in upstream social determinants of health and the public health sector, on the other. Given that people’s health is largely shaped by the daily conditions in which they are born, grow, live, work and age – conditions collectively described as the ‘social determinants of health’ (WHO 2008) – this skewed distribution of resources is simultaneously irrational and unjust, perpetuating racialized, gendered and classed health inequities. While AI and AI4SG projects did not initiate these tensions and inequalities, big data analytics and AI systems are in many ways primed to exacerbate and conceal them. These points provide an entrée to reflecting on the expanding universe of efforts to ‘do good’ through and with AI. In the process, this paper delves into the contradictions, conflicts of interest and harmful consequences – the silences – of AI4SG projects; it highlights social and political issues that such tech-centric projects tend to disregard. In closing, it discusses alternative possibilities to foster progressive social change and democratic renewal – whether alongside, or in spite of, existing and emerging AI technologies.

AI4SG silences: private interests, technological harms and social foreclosures

Both AI and AI4SG are ambiguous terms. Though references to AI and deployments of computer automation in everyday life are ubiquitous – from responding to phone calls to processing drive-thru orders to tailoring online marketing – AI eludes straightforward definition; its meanings are multiple and

historically contingent, depending on, among other things, contemporary technological affordances, social agendas and interpretations, and corresponding parameters of machine ‘intelligence’. For instance, according to Colin Garvey (2018), one crucial shift in AI history took place between the waning of the ‘expert systems’ paradigm (late 1970s–80s) and the rise of the ‘machine learning’ paradigm (2010–present). Expert systems AI, focused on programming specific human expertise and decision-making (like medical diagnostics) into computers, though more successful than earlier quests to develop general AI, was limited by ‘knowledge acquisition bottlenecks’ in tapping and translating the complexities of experts’ knowledge to computers (Garvey 2018). However, the growth of digital big data streams and increased computational power enabled a new AI paradigm that sidestepped such bottlenecks: machine learning AI, through which computers register or ‘learn’ patterns directly from massive data sets, on which algorithms are said to be trained (e.g. to identify images of ‘bicycles’ or ‘angry’ facial expressions). Such technical processes may be the basis for both AI4SG projects as well as AI and big data operations in society more generally. While a fuller discussion of AI techniques and their nuances is beyond the scope of this paper, there is evidence that machine learning capabilities are increasingly prominent in AI4SG efforts across domains, especially in healthcare (Shi, Wang, and Fang 2020, 7).

That said, people may also invoke AI’s cutting-edge aura to hype up computerized data analysis projects, which, upon closer examination, involve nothing more than conventional statistical analyses of large data sets; yet such big data analytics may be included in AI4SG landscape analyses (Chui et al. 2018, 2). Though these large-scale data analyses may be impressive, such statistical calculations have little in common with the machine-based learning processes and neural network algorithms that many currently believe define AI. Nevertheless, such stories can exploit the ambiguous, emergent, speculative nature of AI to build up the AI4SG brand, and the branders themselves, opening up new funding streams in the process. The lines blur between big data analytics, AI, AI4SG and supposed breakthroughs, across fields of science and technology (Cooper and Paneth 2020a, 2020b; Fortun 2008), with some analysts observing, ‘Online tech-hyping articles are now driven by the same dynamics as fake news’ (Funk 2019). These stories also participate in the broader, persistent mythology of technological solutionism and tech-fixes for social and political problems (Huesemann and Huesemann 2011), including emerging variants of computational social science in which researchers claim that new big data streams ‘will make it possible to learn far more about society and to eventually start solving – actually solving – the major problems that affect the well-being of human populations’ (Ledford 2020, 329).

The nascent AI4SG field is embedded in these technical and cultural contexts, led by many of the same corporate actors that are incubating AI systems broadly. Stakeholders shaping this landscape include: (1) Big Tech

companies, including Amazon, Apple, Facebook, Google, IBM and Microsoft, which have formed a Partnership on AI to Benefit People and Society (<https://www.partnershiponai.org/>), beginning in 2016; (2) corporate management consultancies (e.g. McKinsey, Deloitte, Accenture), which frame the field in strategic, opportunistic terms overlapping with older discourses of e.g. corporate social responsibility; (3) established and new NGOs that seek to deploy AI technologies in their work (from groups that may use AI as one of their many activities, such as Amnesty International, to NGOs founded around the use of AI in a specific context, such as Talking Points); and (4) the UN, which since 2017 has hosted an annual AI for Good Global Summit (<https://aiforgood.itu.int/>), oriented toward addressing its 17 Sustainable Development Goals (SDGs). Hence the UN's SDGs have become one influential reference point for many AI4SG projects (e.g. Google 2019; Chui et al. 2018). Numerous academic research centres aspire to contribute to the AI4SG field as well. Government policy roadmaps also use AI4SG language and frames, highlighting ways that AI can contribute to 'the public good' and to addressing 'grand challenges'. Meanwhile scholars have found such reports, touting values such as transparency, accountability and 'positive impact', to be short on details, long-term strategies and meaningful commitments to achieve such impacts (Cath et al. 2017).

Whatever their intentions, such language and frames are not only vague; they foreground a presumed 'social good' while directing attention away from the cultural, economic and political power commanded by Big Tech companies and the harmful externalities inherent in burgeoning AI systems. Indeed, AI and AI4SG are in many ways constituted by and entrench colonial relations (Adams, this issue; Mohamed, Png, and Isaac 2020), characterized by hierarchy, extraction, exploitation and malevolent paternalism (Amrute 2019), within and across nation-state borders. For example, in an era of worsening climate change, researchers have drawn attention to the enormous quantities of energy required to train algorithms on big data sets, leading to huge carbon footprints (Strubell, Ganesh, and McCallum 2019), with consequences that will most severely affect the most vulnerable communities. In addition, scholars have analyzed how training algorithms entails extensive low-wage work, including to clean and label the underlying data and to monitor on-going algorithmic operations for accuracy, in conditions that are frequently exploitative, particularly in the Global South (Gray and Suri 2019). For these reasons, among others, Jared Moore has proposed reorienting AI4SG toward 'AI for Not Bad', to more accurately characterize the complexities at stake:

'AI for social good' speaks to the desire of many of practitioners to share what opportunities they have. It sounds nice. It imagines a world of lucrative careers optimized to better humanity. The world is not so simple ... AI practitioners, like myself, are part of the prospecting of science from which we hope for gold, but in which we will likely find just sand – and perhaps leave in our tailings environmental damage and labor

displacement. Lest that be so, we must be honest about what we are doing and what we might do better' (Moore 2019, 6).

However, these reflections and their underlying empirical grounding are a far cry from the dominant refrains of AI4SG practitioners and advocates (Google AI Impact Challenge 2019), even as AI4SG initiatives may increasingly sound notes of greater skepticism, precaution and reflexivity (Tekisalp 2020) in response to such critiques (Latonero 2019). More significant are the issues typically left off the AI4SG table entirely: AI's own ecological footprints; AI's racialized, gendered labour inequalities; the broader social costs of AI's labour market 'disruptions', including not only labour displacement but greater labour surveillance; and meaningful government regulation and democratic governance of AI R&D, including Big Tech stakeholders – to name a few. Ultimately these silences foster foreclosed social possibilities, with US policies 'distorted against labour and in favour of capital' (Acemoglu, Manera, and Restrep 2020), even as the growth of the AI4SG field accelerates.

Big data, AI and public health

Currently, health and medicine are a leading application sector for big data analytics, machine learning and AI, with health care 'the largest market for investment in the emerging AI/ML business sector' (Ostherr 2019; the underlying business analytics data are proprietary, however). This popularity includes AI4SG projects, which, as noted in the introduction, are often synonymous with commercial AI activities in general, in settings from Big Tech companies to start-ups (though perhaps rebranded as 'social enterprises'). Google's 2019 report on its first Google AI Impact Challenge, which analyzed the AI4SG landscape revealed by the 2,602 project proposals it received from organizations in 119 countries (of which 20 received funding and capacity-building support), found that, 'Of the sectors represented, health-related applications were the most common, representing more than 25% of total submissions' (Google 2019, 7). A 2018 McKinsey AI4SG white paper also found the greatest number of actual or potential use cases in its 'Health and hunger' domain, comprising 28 of 160 use cases (Chui et al. 2018). Meanwhile the AI4SG literature survey mentioned above found more papers in the 'Healthcare' sector (comprising both clinical care and public health) than in any of the eight application domains it identified (Shi, Wang, and Fang 2020, 5–6), with this sector accounting for 32% of the 2019 AI4SG literature surveyed, as well as having the highest rate of growth in recent years, such that 'the difference between it and other domains appears to be widening' (Shi, Wang, and Fang 2020, 5). Articulating their enthusiasm for leveraging big data for public health, Muin Khoury and John Ioannidis write, 'Big Data stands to improve health by providing insights into the causes and outcomes of disease, better drug targets for precision medicine, and enhanced disease prediction and prevention' (Khoury and Ioannidis

2014, 1054). AI and big data work hand in hand in such a project – an interplay of technology, analysis and mythology (boyd and Crawford 2012) – identifying patterns across datasets from genomics to electronic health records (EHRs) to infectious disease outbreak data to social media and financial data.

While on the face of it these trends and new technological affordances may seem to be good news for those concerned about health as a social good, such AI4SG and AI interventions tend to miss the mark in multiple ways when it comes to promoting public health and health equity, rooted in social justice (Krieger and Birn 1998). Potential issues include: (1) distracting attention and resources away from root causes of population health inequities and upstream social determinants of health; (2) enabling new risks and vulnerabilities, especially for already marginalized communities; (3) failing to adequately grapple with the politics of data, knowledge and expertise in the ‘computational turn’ to big data analytics and AI; and (4) fostering new data treadmills, rather than emphasizing existing warrants for action based on current bodies of research. The points below unpack some of these issues and their implications for public health, so-called AI4SG projects and AI governance, with an eye to health equity.

1) Distracting from root causes and upstream social determinants of health

First, the emphasis on the novel possibilities of big data and AI for public health, in AI4SG projects and more generally, often dovetails with a particular, narrow, empirically inadequate framing of health: the so-called medical model (Iton 2010), which has long been in tension with more capacious, social-ecological frameworks oriented population health promotion, prevention, social justice and health equity (Brandt and Gardner 2000; Fairchild et al. 2010). The medical model, conceptually and practically, emphasizes downstream, individualized interventions in health care settings, increasingly wedded to emerging precision medicine treatments, for which big data and AI are enabling technologies (Ho et al. 2020; Hager et al. 2017; Schork 2019; Sun et al. 2018). As noted in the AI4SG literature survey, ‘The majority of research efforts in AI for healthcare are focused on clinical care’, citing the availability of EHRs and medical images as having ‘directly facilitated the AI research in disease diagnosis, clinical treatment, and clinical prediction’ (Shi, Wang, and Fang 2020, 20). Among these applications, ‘Disease diagnosis is the topic in healthcare that has seen the most applications of AI’ (again, with this survey including both clinical care and public health in its ‘healthcare’ category; (Ibid.)). For example, there are now AI-based smartphone apps for diagnoses of myriad health issues, from chronic obstructive pulmonary disease to pneumonia and acute asthma (Stewart 2017) to skin cancer (Comstock 2018; Chui et al. 2018) to inflammatory skin diseases (Wu, Yin, and Chen 2020) and beyond, in addition to the

many apps for monitoring and managing health issues, such as heart disease (Harvard Heart Letter 2019). The Nuffield Council on Bioethics has also highlighted the preponderance of clinical biomedical applications of AI in health care and research (2018, 3-4). Meanwhile the need for not only medical diagnosis but also access to care may be referred to in the AI4SG field as a ‘last mile problem’ (Chui et al. 2018, 20).

What is not explicitly discussed in this AI4SG survey, nor in other discussions that centre AI and big data as means of advancing health, are the politics of different types of health interventions: in particular, in a US context, the tension between growing investments in ever more individualized treatments in downstream clinical settings, on the one hand, and persistent underinvestment in upstream social determinants of health and the public health sector, on the other (Bradley and Taylor 2013; Interlandi 2020). Yet as one recent article put it: ‘An undeclared civil war is breaking out in biomedicine. On the one side is precision medicine, with its emphasis on tailoring treatments to ever-narrower groups of patients. On the other side is population health, which emphasizes predominantly preventive interventions that have broad applications across populations’ (Cooper and Paneth 2020b). For while precision medicine privileges individual-level variables – especially genetics, in practice – it deemphasizes the everyday environments and economic and social resources (e.g. adequate income, affordable housing, access to healthy food, freedom from discrimination, clean air and water, accessible parks, high-quality education, health insurance) that are most consequential to health (Braveman and Gottlieb 2014; Galea 2019; WHO 2008). It also neglects the political and social structures that shape the distribution of these resources (Beckfield 2018). In the process, precision medicine decentres structural racism and exploitative economic relations, intertwined with people’s diverse intersectional identities, as root causes of public health inequities (c.f. Figure 1), as illuminated by multi-level ecosocial frameworks in contemporary social epidemiology (Krieger 2011, 214, 287). As physician and public health leader Anthony Iton puts it, ‘Common disease roots in the socioecological context are often ignored [in the medical model]’ (Iton 2010, 512). In addition, the medical model of health has often been accompanied by unscientific, false conceptions of biological race in clinical research and practice, including by positing ‘race’ – rather than racism – as a cause of racialized health inequities (Boyd et al. 2020; Chadha et al. 2020).

While AI and AI4SG projects did not create these issues and inequalities, big data analytics and AI systems are in many ways primed to exacerbate and conceal them. For example, a range of scholars have highlighted the problems with the US’s disproportionate, growing investment in an individually focused, clinically oriented, AI-enabled precision medicine agenda versus a population health and health equity agenda, focused on investments in upstream social determinants of health (Bayer and Galea 2015; Cooper and Paneth 2020a;

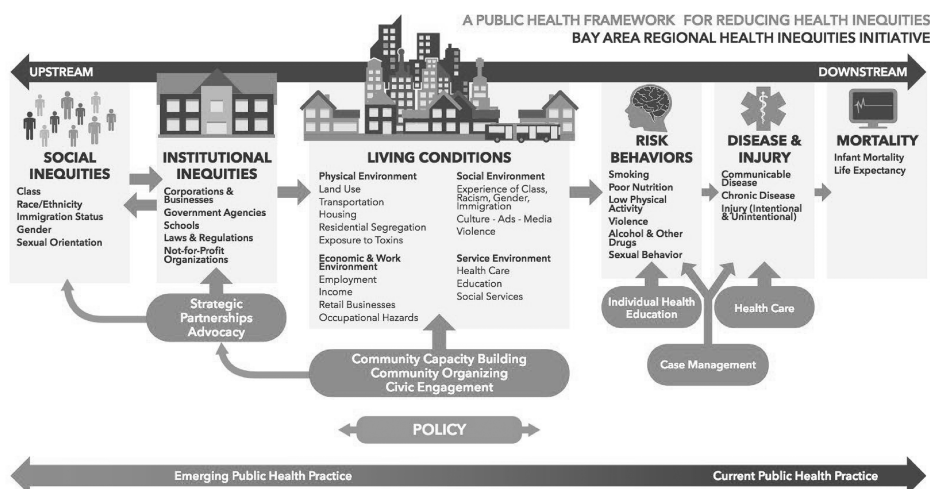


Figure 1. A Public Health Framework for Reducing Health Inequities. Source: Bay Area Regional Health Inequities Initiative, as cited in 'Portrait of Promise: The California Statewide Plan to Promote Health and Mental Health Equity', California Department of Public Health (CDPH), Office of Health Equity (2015), 17.

Chowkwanyun, Bayer, and Galea 2018; Ferryman and Pitcan 2018). Ferryman and Pitcan's 'Fairness in Precision Medicine Study' (2018), assessing equity issues raised by precision medicine, concludes that not only is there potential for biased datasets, as big data advocates acknowledge, there is also the rarely addressed problem of biased outcomes due to diversion of resources away from structural, social determinants of health (Ferryman and Winn 2018, 34). This point speaks to longstanding skews in US health investments and outcomes, compared with other high-income countries (Kristof 2020; Woolf and Aron 2013). However, as scholars Bradley and Taylor document in *The American Healthcare Paradox* (2013), the purported paradox – of high per capita US healthcare spending, coupled with worse US population health outcomes – disappears once the US's much lower spending on social services and safety net programs is taken into account. That is, rather than promoting health for all through upstream access to resources for flourishing, the USA chooses to medicalize the health effects of racialized, gendered, classed deprivation downstream, through the health care system (and to provide social services that are not only inadequate, but often punitive and disempowering; c.f. Hatton 2020). Another recent report (Lown Institute 2019) highlights these same patterns at the state level in California, a hotbed of AI innovation and AI4SG; it finds that California spends only \$0.68 on social services, public health and environmental protection for each \$1.00 spent on health care – and that these disparities have worsened since 2007. Moreover, 'with increasing sophistication of medical technology, the American tendency to medicalize patient concerns is widely recognized as contributing to health care cost escalation' (Bradley and Taylor 2013, 162), with enormous implications for health

equity (Woolf et al. 2008), even as algorithms can further embed racial discrimination in medicine (Obermeyer et al. 2019). Hence, both the financial and the opportunity costs of AI technologies in health care – often oriented toward individualized health diagnostics, monitoring and ‘precision’ medical treatments in downstream health care settings – are prone to distract attention and resources away from root causes of health inequities and upstream social determinants of health.

That said, public health discourses, emphasizing social determinants of health, are increasingly being taken up in conjunction with AI-based interventions, particularly in health care settings, sometimes intertwined with precision medicine (for a fuller critical discussion of ‘precision public health’, c.f. Chowkanyun, Bayer, and Galea 2018, 2019). Advocates point to the potential of AI to better identify ‘high-risk’ patients, based on social data (e.g. people at risk because they lack access to basic resources in their daily lives, often long prior to becoming patients for particular health issues), to incorporate these social determinants of health data into individual EHRs and clinical workflows, and to refer patients to social services and monitor the results (Cognizant 2019). Such possibilities, which have opened up in part due to new ‘value-based care’ payment models that emphasize social determinants of health for the first time (Cognizant 2019, 5; Ostherr 2019), may help to mitigate inequitable health care outcomes among patients, especially if the inadequacies of current value-based models vis-à-vis reducing racialized and other health inequities are addressed (Ojo, Erfani, and Shah 2020).

However, action on social determinants of health at the relatively downstream, health care level (e.g. by providers and insurers), including by using AI to connect people with non-medical social services and to monitor outcomes, does not address the larger picture of skewed health investments discussed above (Alberti, Bonham, and Kirch 2013; Maani and Galea 2020a; Silverstein, Hsu, and Bell 2019). For despite the promise some see in myriad AI and AI4SG health use cases, ‘Achievement of primary prevention benefits depends more on social factors than secondary prevention irrespective of the marginal benefits of artificial intelligence’ (Panch et al. 2019, e13) – upstream social factors like jobs and shared prosperity that AI may, on balance, be disrupting in ways more harmful than not. As medical and public health scholars Nason Maani and Sandro Galea write, ‘[T]he notion of addressing social determinants in the context of clinical practice devalues and medicalizes the complex burden and barriers encountered by those affected by discrimination or poverty’ (Maani and Galea 2020a, E1). Instead of new data analytics, they underscore the centrality of distributions of power: ‘Many fundamental determinants of health are far upstream of health care and are deeply rooted in the distribution of money and power, at local and national levels’ (Maani and Galea 2020a, E1). And as other scholars point out, ‘[F]orce-fitting strategies to address social determinants of health into traditional models of clinical care risks

misdirecting limited resources into programs that may ultimately prove inefficient or ineffective' (Silverstein, Hsu, and Bell 2019, E2).

In addition, the potential for commercial entities, including private Big Tech companies, to coopt and attempt to 'redefine' – and monetize – public health in these contexts is deeply troubling, including claims by Facebook representatives about the promise of using social media and online consumer data to: 'transform the traditionally held social determinants of health, including education, income, housing and community, to encompass a more granular tech-influenced definition, ranging from simple factors, such as numbers of online friends, to complex social biomarkers, such as timing, frequency, content and patterns of posts and degree of integration with online communities' (Abnoui, Rumsfeld, and Krumholz 2019, 247). Again, tech-centrism and tech-solutionism can distract from seeing the inequalities in power and resources that drive population health inequities, as well as the relative lack of accountability of health care, pharmaceutical and technology companies to the public health field and diverse publics. While this paper focuses on these issues in a US context, the power of the US tech sector and tech philanthropies in shaping public health is certainly global in scope (Birn 2014), often perpetuating colonial relations in the process (Amrute 2019).

2) Big data and AI: new vulnerabilities and risks

Second, not only are AI and AI4SG projects prone to misdirect attention and resources away from root causes of social and environmental justice issues, including public health; they also enable new forms of discrimination, disadvantage and vulnerability resulting from the confluence of emerging big data streams, machine learning algorithms and increasingly personalized 'risk' profiles (Achiume 2020; Figueroa, Frakt, and Jha 2020; Mittelstadt and Floridi 2016). These new vulnerabilities and sources of discrimination are primed to multiply in the years ahead, particularly as Big Tech companies and other profit-oriented entities mine people's online data to sort and stratify them (Abnoui, Rumsfeld, and Krumholz 2019), and data are combined and used in ways people never imagined, much less consented to. Even advocates of proposed 'polysocial risk scores', which would link multiple social data streams to profile people in health contexts, acknowledge, '[T]he risk of false-positives and false-negatives could be high, especially in the early iterations' (Figueroa, Frakt, and Jha 2020, E2).

The risks posed by new risk scores and profiling will tend to deepen the challenges faced by already marginalized communities, potentially worsening racial and other forms of discrimination through algorithmic decision-making (Obermeyer et al. 2019; Heaven 2020), not only at the individual level but at the level of social groups and collectivities, in both the private and public sectors (Engstrom et al. 2020). As Ferryman and Pitcan quote bioethicist Lisa

Park as saying, in an interview regarding precision medicine data analytics: ‘[I]f there are particular health risks that are associated with traditionally underrepresented, underserved, or discriminated-against populations, there’s an opportunity or a likelihood that there will be exacerbation of that discrimination’ (Ferryman and Pitcan 2018, 30). Moreover, Park notes that information about people’s purported health risks could be used coercively by a range of public and private actors – for example, to condition people’s access to social services or insurance coverage. The result will be that less enfranchised people who are already exposed to extensive surveillance and constraints on their autonomy will encounter additional vulnerabilities to repressive risk assessments, controls and sanctions. If automated, through the use of algorithms, these new risk assessments could both magnify existing social systems of exclusion and oppression while also being more opaque and difficult to challenge, as ‘Jim Crow practices feed the ‘New Jim Code’ – automated systems that hide, speed and deepen racial discrimination behind a veneer of technical neutrality’ (Benjamin 2019b, 422; also c.f. Benjamin 2019a). These current and anticipated risks point to the inadequacies of conventional, individually oriented bioethics frameworks in grappling with emerging technologies (Obasogie and Darnovsky 2018), including as a touchstone for computer and AI ethics (Floridi et al. 2018). They point to the need for a ‘new biopolitics’ (Obasogie and Darnovsky 2018), or as technology scholar Sareeta Amrute calls for, ‘Radicalize and *politicize* your ethics – critique, reimagine, practice, repeat’ (Amrute 2019).

Scholars and activists collaborating on the Our Data Bodies Project (<https://www.odbproject.org/>) likewise illuminate these politics in their work, analyzing the new risks and vulnerabilities enabled by big data analytics and AI. Among the findings of their *Reclaiming Our Data* report (2018), which draws on interviews with residents of Charlotte, Detroit and Los Angeles: (1) ‘Predatory data-driven systems routinely hold us back and prevent us from meeting basic needs’, e.g. through the use of credit scores in job and housing applications, or otherwise subjecting people to extra stigma, scrutiny or punishment by social institutions – from law enforcement to health care to public services – as well as to additional uncertainties, depending on how and by whom data might be used in the future; (2) ‘Data-driven systems unfairly divert us from resources we are entitled to and need to survive’, e.g. through one-way, asymmetric data collection by powerful government and corporate actors, even as people’s basic needs and information that might be helpful to meet them are neglected, thereby reinforcing existing power relationships; (3) ‘The hurt and the harms of data-driven systems connect to past systems of discrimination and exploitation against members of marginalized communities’, e.g. when people are profiled in terms of race or ethnicity, socio-economic status, age, gender, sexuality, housing status, criminal record, or some combination; and (4) ‘Our policies, priorities and visions for the future focus on just models of

governance, economics and criminal justice, and on guarantees for our human rights', e.g. by emphasizing inclusion of all, sharing and empathy, as well as renewing 'off the grid' connections with people and places as a strategy for self-defense and survival (Petty et al. 2018; also c.f. Lewis et al. 2018).

These insights, informed by deeper recognition and analysis of social inequities impinging on people's well-being, including structural racism, bring a vital corrective to over-hyped prognostications about leveraging big data and AI on behalf of public health and other social goods. Indeed, they highlight the potentials for automating inequalities and 'feedback loops of injustice' (Eubanks 2018), rather than reducing or dismantling them. Such considerations are all the more important given that many big data and AI applications are developed in private corporate or start-up settings, beyond democratic governance, and are deployed for profit (Singer 2017). Companies like Facebook have already demonstrated negligence in safeguarding people's personal data, which in health care contexts could lead to people being denied coverage by an insurer, employment discrimination or other harms (Singer 2019). These conjunctures point to the urgency of innovations such as Data for Black Lives' proposed Public Data Trust (Milner 2018). Moreover, attending to biases in algorithms and training data is crucial and necessary, but not sufficient, to achieving social justice, when entire social systems are unjust – and given that 'algorithmic fairness' is never a merely technical, objective matter (Heaven 2020; Noble 2018).

3) The politics of data and expertise in the computational turn

Third, discussions of big data, AI and public health (among other 'social good' domains) often fail to adequately grapple with the politics of data, knowledge and expertise – in academia and beyond – in the so-called computational turn. As boyd and Crawford note, 'Big Data changes the definition of knowledge', as it refers not only to vast quantities of data and procedures for analyzing them, but 'also to a computational turn in thought and research' (boyd and Crawford 2012, 665), as in computational social science, wherein new forms of digital, big data-mediated research (e.g. computerized social network or social media analysis) are framed as the cutting edge of insights into social life. This computational bias increasingly shapes the very conceptions of data and research embedded in nominally interdisciplinary projects and data infrastructures, perhaps especially those framed around open science, which emphasizes digital data and its 'usability' (including machine-readability) across fields. As the International Science Council's Committee on Data for Science and Technology (CODATA) describes, it aims to 'promote global collaboration to advance Open Science and to improve the availability and usability of data for all areas of research', while defining data as, 'Facts, measurements, recordings, records, or observations about the world collected by scientists and others,

with a minimum of contextual interpretation' (<https://codata.org>). That is, even while ostensibly devoted to data in all areas of research, CODATA's conception of data actually presumes a narrow, positivist approach to knowledge and data that excludes and implicitly devalues interpretive, contextually attuned research – whether in the humanities, social sciences, community-based participatory research, or other fields. A synergy among big datasets, machine learning and open science rhetorics seems likely to further entrench these computational, positivist biases in research and knowledge. That said, the Research Data Alliance, organized by international stakeholders with a similar mission to 'develop and adopt infrastructure that promotes data-sharing and data-driven research' (<https://www.rd-alliance.org/>), articulates a more reflexive approach to interdisciplinarity (Lélé and Norgaard 2005), better attuned to disciplinary specificities and contexts of 'data', indicating the wider array of possible institutions to navigate these issues (e.g. the Platform for Experimental, Collaborative Ethnography: <https://worldpece.org/>). Though in many cases, open sharing of data and research may not be desired by or serve communities (Albornoz 2018), as a range of activists and scholars, including founders of the Global Indigenous Data Alliance (<https://www.gida-global.org>), have been leaders in establishing.

At the same time, some big data advocates propose that the challenges of triangulating diverse data streams – of intra – and interdisciplinary knowledge syntheses – might be overcome by algorithms. Regarding big data and public health, Khoury and Ioannidis assert that, 'There also must be a means to integrate knowledge that is based on a highly iterative process of interpreting what we know and don't know from within and across scientific disciplines' (Khoury and Ioannidis 2014, 1054). They then suggest that machine learning algorithms might be used to such ends, citing the example of the ClinGen project to 'create centralized resources of clinically annotated genes to improve interpretation of genomic variation and optimize the use of genomics in practice' (Khoury and Ioannidis 2014). Meanwhile they neglect the broader public health field and the politics and challenges of interdisciplinarity across fields such as computational genomics and social epidemiology, or a wider range of social and natural sciences, from political ecology to social psychology. These fields may be centred on disparate units and scales of analysis, methodologies, evidentiary standards, epistemologies, values and more – rendering their respective bodies of knowledge in tension with one another (or even incommensurable, from some vantages; c.f. Espeland 1998). These tensions belie efforts to seamlessly synthesize and integrate knowledge, as more critical scholarship on interdisciplinarity has highlighted (Brown, Morello-Frosch, and Zavestoski 2011; Lélé and Norgaard 2005; Sarewitz 2004, 2015; Strober 2010). As social epidemiologists Nancy Krieger and George Davey Smith write, '[D]ata never speak by themselves – either to computer algorithms or to people – and nor do beliefs about probabilities simply drop from the sky. Active scientific judgment

is inevitably involved ...'. (Krieger and Smith 2016, 1794). Algorithms claimed by some to be innovative, neutral and objective as means to 'integrate knowledge' and define 'evidence-based' practice instead conceal persistent epistemic politics and hierarchies (Braveman et al. 2011).

More broadly, inclusive and critical knowledge syntheses would not be limited to the research and data of credentialed experts, from whatever disciplinary fields, or to digital data. Instead, they would incorporate street science (Corburn 2005), popular epidemiology and lay health expertise (Brown 1992; Epstein 1995), indigenous ecological knowledge (Anderson 2005), popular education (Onuoha 2018), and other forms of community-based, embodied, grassroots expertise, including to contest and reframe conventional approaches to data, which may otherwise institutionalize racial inequities and other biases (Hawn Nelson et al. 2020). In the public health field, such knowledge syntheses are all the more crucial when they intertwine with the politics of diagnoses and 'contested illnesses', or 'conditions whose causes are either unexplained by current medical knowledge or whose purported environmental explanations are in dispute' (Brown, Morello-Frosch, and Zavestoski 2011, 18), from asthma to cancer to Gulf War-related illnesses. Environmental justice and other health social movements may conduct community-based research, both independently and in collaboration with academic researchers, to advance new bodies of knowledge that synthesize diverse forms of expertise in confronting these illnesses (Corburn 2005). In the process, health social movements aim to transform conventional hierarchies of credibility while also directing attention to – and advocating for change around – environmental and other dimensions of disease that are often relatively neglected (Brown, Morello-Frosch, and Zavestoski 2011).

Again, the politics of expertise and science are bound up with the politics of different kinds of interventions and social transformations (Martin 2006), or lack thereof. While physicians may be concerned with the status of their expertise vis-à-vis AI technologies in clinical settings, as discussed elsewhere in this issue (Hanemaayer; Burton et al., this issue), expertise pertaining to (or impinging on) health equity extends far beyond clinics; it encompasses the public health field as well as broader political, economic and social policy-making, as recognized in 'Health in All Policies' frameworks (Rudolph et al. 2013). Building trust and investment in AI at one level, such as in clinical care, may mean further entrenching hierarchies and inequalities that diminish health equity overall (Pūras 2020), even if some patients receive high-quality care. This wide spectrum of decision-making, knowledge and values begs the question of which and whose 'social good' – or conception of 'optimization' – AI and AI4SG projects purport to advance. In any given big data or AI project, there will be a politics of 'domain expertise'; some people's values, priorities, epistemologies, types of evidence, forms of data and diagnostic or prescriptive frames will predominate – marginalizing or foreclosing other possibilities,

especially from systemic perspectives. Machine learning algorithms cannot escape or transcend these contentious, value-laden issues. Indeed, to date, widely used search engine algorithms have instead been found to reinforce racism and sexism in people's everyday digital lives, through the types of 'discoverability' they foster and suppress (Noble 2018).

4) New data treadmills, or new commitments to action on existing research?

Fourth, AI4SG and AI discussions of public health and other social issues often neglect warrants for action based on existing research. Advocates for leveraging big data on behalf of health may not even be familiar with voluminous bodies of existing public health research, much less the politics and policies impinging on action on that research (Hofrichter and Bhatia 2010), which their efforts are further shaping. Yet overweening attention to big data and AI projects could displace the value and transformative potential of current knowledge – from public health and social science research to local knowledge and environmental health documentation (e.g. activist reports, oral histories, journalism, photography). Big data and AI could also lead to new cycles of (re)validating that environmental and social variables matter to health, not only genes and individual biomarkers – fostering new promissory 'precision' data horizons (Kuch, Kearnes, and Gulson 2020), 'data treadmills' (Wylie, Shapiro, and Liboiron 2017, 412), the 'datafication of injustice' (Benjamin 2019a, 116) and the continued rationalization of inaction. As Benjamin writes, 'Demanding more data on subjects that we already know much about is, in my estimation, a perversion of knowledge' (Benjamin 2019a, 116). Yet big data, AI and AI4SG advocates tend to highlight only the opportunities presented by new data streams, and the supposed inadequacy of current data (Abnoui, Rumsfeld, and Krumholz 2019; Figueroa, Frakt, and Jha 2020), rather than the myriad unrealized opportunities for policy-makers to call for action on current bodies of research.

As one example, what if US scientific, health and business leaders were equally or more enthusiastic about acting on current research pertaining to healthy early childhood development, including healthy neurological development, as they are about precision medicine and genomics? What if AI4SG advocates who are so concerned about the 'opportunity costs' of underutilizing AI (Floridi et al. 2018) reckoned instead with the opportunity costs of perpetual underinvestment in social determinants of health? What if, instead of heralding the promise of new big data streams and AI, they resolved to mobilize resources to act on the findings of past National Academies research syntheses on these topics, such as *From Neurons to Neighborhoods: The Science of Early Childhood Development* (Shonkoff and Phillips 2000) and *Vibrant and Healthy Kids: Aligning Science, Practice, and Policy to Advance Health Equity* (DeVoe, Geller, and Negussie 2019)? These reports highlight the science of Adverse

Childhood Experiences (ACEs), toxic stress, and the importance of social determinants of health to children's healthy development. Yet decision-makers at multiple levels have by no means prioritized or adequately acted on the findings – resulting in a chasm between knowledge and practice that more big data and AI will not resolve. Nor do efforts by some precision medicine advocates (California Precision Medicine Advisory Committee 2019) to incorporate 'social determinants of health' and ACEs data in relatively downstream, individualized clinical care adequately address these issues, as discussed previously.

Instead, as Wallack and Thornburg indicate in their analysis of intervening upstream for healthy childhood development (Wallack and Thornburg 2016), there is a need to recognize the centrality of politics and new political formations to innovation for health equity, rather than merely more data and research. After reiterating an observation from Thomas Pynchon's *Gravity's Rainbow* that, 'If they can get you asking the wrong questions, the answers don't matter', they go on to suggest the following 'right questions', emphasizing the political and social contexts that invariably mediate research translation into practice and everyday life:

'If any particular geographic area or region were to become the healthiest place in the world to be pregnant and have a child, what would it look like? ... What kinds of policies would be required to move toward that vision? How can we create a social movement built on this collective vision to force the necessary political will to demand change? How can existing partnerships be expanded? How can we develop new partnerships with new allies to move ahead? What political barriers must we overcome?' (Wallack and Thornburg 2016, 938).

After witnessing years of inaction on the science of early childhood development and social determinants of health, Wallack and Thornburg do not emphasize the potential for new research and data streams to suddenly catalyze breakthrough interventions. Rather, they articulate a broader theory of change (and call for action, in policy and practice), encompassing social and political contexts of data and sociotechnical systems. In contrast, many AI4SG advocates and projects lack such broad visions and theories of change, rooted in existing and emerging social movements; nor do they often designate grassroots activists and social movement leaders as AI4SG 'domain experts,' thought partners and white paper audiences (e.g. Rolnick, Donti, and Kaack 2019), even when disavowing a belief in tech-fixes. AI4SG initiatives therefore often fail to critically discuss technical and political dimensions of interventions, perpetuating technocratic decision-making, despite voluminous scholarship in these areas (Asdal, Brenna, and Moser 2007; Ferguson 1994), including work to develop greater 'structural competency' among health professionals, 're-centring the political elements of disease distribution' (Neff 2020, 8; Metzl and Hansen 2014). Such analyses are all the more crucial

when extensive research and data are available to warrant and guide action, yet confront deep resistance in some quarters, including denials of structural racism. As another recent analysis put it, '[M]any agency solutions and data initiatives are largely disconnected from this root cause [of structural racism], and the 'hunt for more data is [often] a barrier for acting on what we already know' (Benjamin 2019a)' (Hawn Nelson et al. 2020). While more investment in the public health sector's data infrastructures could support broader public health efforts, including in response to the current Covid-19 pandemic, deeper investments in upstream social determinants of health for all – not just more data collection – are ultimately called for. Again, while AI and AI4SG projects did not create these issues and inequalities, AI systems are poised to exacerbate and conceal them in multiple ways, given the vast inequalities, distributions of power and social injustices of the current world. These interrelated issues point to common blindspots and cross-cutting biases in the larger universe of AI4SG initiatives, as discussed further below.

AI for not bad and social transformations for health equity

In sum, the confluence of big data, AI and public health presents multiple challenges from a social justice vantage, oriented toward health equity for all. These challenges include (though are not limited to): (1) distracting attention and resources away from root causes of health inequities and social determinants of health; (2) enabling new risks and vulnerabilities due to emerging data streams, especially for already marginalized communities; (3) failing to adequately grapple with the politics of data, knowledge and expertise in the 'computational turn' to big data analytics and AI; and (4) fostering new data treadmills, rather than emphasizing existing warrants for action based on current bodies of research. AI4SG initiatives often pre-frame systemic social and environmental problems in tech-centric ways, while suggesting that addressing such problems hinges on more or better data. They thereby perpetuate incomplete, distorted models of social change that claim to be 'data-driven'. In the process, AI4SG initiatives may obscure or 'ethics wash' all the other uses of big data analytics and AI that more routinely serve private interests and exacerbate social inequalities – including through AI's ecological footprints, displacement and exploitation of workers, surveillance of publics, heightened discrimination against already marginalized groups, and inadequate social benefits from publicly funded R&D. So, however well-intended, a more appropriate frame and goal than AI4SG might be 'AI for Not Bad', as Moore (2019) suggested, in keeping with the bioethical principle to 'first, do no harm'.

While AI4SG proponents acknowledge some of these challenges and risks – particularly problems caused by biased datasets, inaccurate predictions, algorithmic opacity and illegal or unintentional breaches of privacy (Chui et al.

2018, 35) – their inattention to or blindspots around larger social, political and economic systems often hamper consideration of the fuller range of social justice issues at stake (e.g. such as those discussed in the 2018 *Reclaiming Our Data* report). The lack of diversity among AI engineers likely contributes to these disconnects, as technologists develop tools that disproportionately reflect their own social backgrounds and interests (Ferryman and Pitcan 2018). AI4SG advocates, whether AI developers or otherwise, are also often removed from the everyday political challenges and social contexts of less enfranchised communities, even when their projects emphasize collaborative community partnerships; meanwhile the people involved with such projects on the ground may well be unfamiliar with the phrase AI for Good and its accompanying influencer circuits, from South by Southwest to Davos. Hierarchies of power in workplaces developing AI and AI4SG projects, from academia to corporate settings, may also invalidate or discourage adequate critique by those closest to such projects (Hatton 2020). Though this article has focused especially on public health issues, these concerns are relevant to a wider gamut of social and environmental challenges toward which AI interventions may be directed.

In particular, projects that tout the promise of new digital big data streams, yet lack clear theories of change connecting those data with desired outcomes, can too easily lead all involved to believe that a project will result in meaningful change, regardless of the prospects for such change. These include AI4SG projects centred on new types of monitoring and tracking (e.g. of climate change indicators, health indicators, educational indicators, or other forms of environmental or social surveillance) that neglect possibilities discussed in this paper – of obfuscation, distraction, the creation of new vulnerabilities, invalidation of grassroots knowledge, and delay and disregard for existing warrants for action. In light of these issues, some key questions for those considering AI4SG projects include:

- 1) ***Who is at the table? Where are the grassroots activists and social movements among the ‘domain experts’?*** Which stakeholders are part of the conversation about a potential project? How are these stakeholders representative, or not, of the array of people active around or affected by a particular issue?
- 2) ***What is ‘the problem’ and its history? What are the root causes?*** What are the social, political, economic and cultural contexts surrounding an issue and potential AI4SG project? Do all stakeholders agree on how to define a problem, as well as hypothetical interventions? Do proposed solutions centre technology and technocratic decision-making? Or are social and political change centred? What are the trade-offs or tensions between different

potential interventions and problem-solving strategies (e.g. more or less tech-centric strategies)?

- 3) ***What is the project's vision or theory of change?*** What are the intended outcomes of a potential AI4SG project, and the steps to get there? What would it mean for a particular issue to be mitigated or 'solved'? What would systems change entail? Who, if anyone, would be responsible for acting on data from the project? What other forces could impinge on desired outcomes? What resources and processes are available to respond to situations indicated by the data? How might the data help with community organizing and power-building? What other means of power-building are available to address a given issue?
- 4) ***What are the politics of knowledge and expertise?*** How might data generated by the project amplify, or undermine, other forms of community knowledge and data (particularly people's lived experiences and community-based research of many kinds)? How might data be contested by various parties? How does the project build on existing technological infrastructures and technological access (or not), including equity in data access, analysis, interpretation and legibility?
- 5) ***What could be the unintended consequences of a project?*** What other outcomes are possible? How might an intervention address or obscure root causes? How might it foster new data treadmills? How might it fit into the broader landscape of problem-solving around a particular issue, including resource allocations? Which resources and processes could help empower those most affected, already marginalized, or newly vulnerable, including if data are misused?
- 6) ***How does the project intersect with racial equity?*** How do a project's data engagements dovetail with guidelines for centring racial equity and confronting structural racism throughout data lifecycles, including as articulated in 'A Toolkit for Centring Racial Equity Throughout Data Integration' (Hawn Nelson et al. 2020), focused on civic data use?
- 7) ***How does the project intersect with health equity?*** How does a project's implicit or explicit theory of change intersect with the APHA's statement on the path to health equity? Namely: 'How do we achieve health equity? We value all people equally. We optimize the conditions in which people are born, grow, live, work, learn and age. We work with other sectors to address the factors that influence health, including employment, housing, education, health care, public safety and food access. We name racism as a force in determining how these social determinants are distributed'.
- 8) ***Why should or shouldn't the project be pursued? How could this assessment change?*** Considering the issues above and any other relevant concerns, what could be evidence of a potential project's (in)efficacy, unintended social or political consequences, or other indicators that a

project should or should not be pursued? How might these indicators be revisited periodically?

These lines of inquiry, and many more that could be generated by stakeholders addressing specific issues, are crucial to surfacing potential AI4SG projects' values and visions of change (or lack thereof), including their conceptions of and approaches to whatever 'good' they seek to advance. They are intended to problematize the technological solutionism and tech-fix frameworks that often underpin such projects, addressing but going beyond the rules of thumb for technological fixes developed by Sarewitz and Nelson (2008), as well as lists of AI4SG 'best practices' that make ritualistic disclaimers that AI is not a 'silver bullet' yet fail to seriously incorporate that point into their overall analysis (Floridi et al. 2020, 1773). By shifting attention toward root causes of social and environmental problems (e.g. in systems of oppression and exploitation); new risks and vulnerabilities; the politics of data, knowledge and expertise; and the pitfalls of perpetual data treadmills, these questions seek to provoke deeper reckoning with not only the risks and benefits of potential projects, but alternative paradigms of problem-solving entirely, oriented toward challenging existing systems of power.

Such questions can also help to evaluate and highlight AI4SG projects that may indeed be quite helpful to address particular issues, if developed in collaboration with community partners. Such projects include ActiveRemediation, a machine learning model designed by university-based researchers to predict the location and aid in the removal of water service lines containing lead in Flint, MI, in the absence of adequate public records (Abernethy et al. 2018; Chui et al. 2018, 26). They also include Talking Points, a natural language processing platform developed in a non-profit context that translates multilingual text messages among teachers, parents and students, to 'driv[e] student success in low-income, diverse areas through building strong partnerships across families, schools and communities' (<https://talkingpts.org/about-us/>). These projects, articulating with environmental and social determinants of health, feature readily useable data and clear users in public municipal and educational contexts, directly helping to address specific community challenges while building on existing technological infrastructures, without fostering new data treadmills. That said, in these cases, too, broader systemic interventions remain crucial to achieving environmental justice and health equity, in Flint and beyond, and to the realization of Talking Points' larger stated mission (i.e. to 'driv[e] student success in low-income, diverse areas'). By extension, project funders, including deep-pocketed Big Tech companies and foundations, should not be perceived as adequately 'doing good' simply by assisting with such efforts, however laudable. Rather, they should be held accountable by policy-makers at multiple levels and across sectors (Rudolph et al. 2013) to ensure that all of the political,

economic and social relations they foster are consistent with advancing healthy childhood development and social determinants of health for all.

At present, in the context of COVID-19, US underinvestment in the public health sector, robust social safety nets and upstream social determinants of health – including workers’ rights and protections – has translated into diagnosis and death rates that are high overall as well as deeply racialized and correlated with income, amplifying preexisting health inequities (Maani and Galea 2020b; Serkez 2020). As Anthony Iton writes, underscoring these preventable vulnerabilities at the intersections of race, class and place:

‘COVID-19 is reminding us that in the United States, when it comes to your health, your zip code is more important than your genetic code. Our country manufactures social vulnerability through policy violence. Policy violence is the intentional absence of protective policy in the face of abject need. Policy violence leaves large segments of our society experiencing constant daily stress as they try to navigate a healthy life without health insurance, decent housing, affordable childcare, paid sick leave, or quality education.... The foundation of American policy violence is racism. Scratch the surface of virtually every failed effort at creating universal policies in this country and you’ll find thinly veiled racism at the root’ (Iton 2020).

So from a wider-view lens, instead of foregrounding large-scale data, AI, or other technologies in discussions of innovation and public health, there is a need to centre larger-scale social transformations to address the multifaceted policy violence Iton describes. In addition, Iton and other US social justice leaders have called for new social compacts (Iton 2017), emphasizing interdependency, collective care and decision-makers’ accountability ‘to serve the well-being of people, including the right to health care, food, education, and shelter’ (<https://newsocialcompact.org/>). Along these lines, contemporary social movements have articulated incisive, multifaceted policy platforms centred on racial, economic and climate justice that could frame AI ethics conversations (Movement for Black Lives 2020; National Domestic Workers Alliance 2020; Poor People’s Campaign 2020; Sunrise Movement 2020), rather than discussing AI4SG and the UN SDGs apart from contexts of pervasive neoliberalism and structural racism (e.g. as in Hager et al. 2017; Floridi et al. 2018, 2020), in the USA and beyond. Scholars proposing a decolonial critical approach to AI have likewise called for renewing affective and political communities as one key tactic (Mohamed, Png, and Isaac 2020).

Centring these social transformations would entail contesting current political and economic structures and the forms of ‘innovation’ they produce. In the USA, such transformations are all the more justified given that the public sector often takes on high levels of risk in funding the development of new technologies, yet the public return on these investments tends to be negligible, compared with the rewards captured by the private sector. This results in a skewed ‘risk-reward nexus’, according to economists Williams Lazonick and Mariana Mazzucato (Lazonick and Mazzucato 2013), and, ultimately, a ‘parasitic innovation

ecosystem’ that undermines both innovation and shared prosperity (Mazzucato 2013, 2018). These dynamics reflect deeply rooted biases of conventional US R&D agendas and science and technology policy-making, which disproportionately serve the interests of the most privileged while mostly neglecting social justice and equity issues, in biomedical research and other fields (Bozeman 2020; Woodhouse and Sarewitz 2007). They also reflect tendencies to seek downstream, technological solutions to social and environmental problems rather than preventing or significantly mitigating them upstream; yet in the current era of apparently accelerating ecological and social crises, there is a need for transformed paradigms on multiple fronts (e.g. as articulated in the recent report: *Resilience Before Disaster: The Need to Build Equitable, Community-Driven Social Infrastructure*, Lou et al. 2020; also Interlandi 2020).

On that note, rather than trying to extend technological infrastructures into new domains, computer and data scientists seeking to do good could instead collaborate with social and environmental justice activists to analyze the values and biases embedded in existing data infrastructures and algorithms, in settings from health care (Chen, Szolovits, and Ghassemi 2019; Obermeyer et al. 2019) to law and policing (Heaven 2020), as some are indeed doing. Or, taking a step back, AI4SG advocates could lobby companies for greater transparency and access to algorithms, which are often proprietary, to enable such scrutiny in the first place. Looking to the future of AI, AI4SG advocates could support efforts to diversify AI R&D – such as Black in AI (<https://blackinai.github.io/>) and Women in Machine Learning (<https://wimlworkshop.org/>) – as well as STEM fields broadly. They could support calls for substantive public accountability and public governance of big data, AI, biomedicine and the tech sector in general, including through public data trusts, in solidarity with diverse social movement activists (Bernhardt 2017; Milner 2018) and organizations such as the Algorithmic Justice League (<https://www.ajlunited.org/>), Coalition for Critical Technology (<https://forcriticaltech.github.io/>), Data for Black Lives (<https://d4bl.org/>), Data & Society (<https://datasociety.net/>), Design Justice Network (<https://designjustice.org/>), Science for the People (<https://scienceforthepeople.org/>), Structural Competency (<https://structuralcompetency.org/>), the Tech Workers Coalition (<https://techworkerscoalition.org/>) and many others concerned with social justice and existing and emerging technologies.

Computer and data scientists could also support health equity advocacy and action on existing public health research, through environmental and social justice organizations as well as the formal public health field (and organizations spanning these sectors, such as Public Health Awakened: <https://publichealthawakened.org/> and the Spirit of 1848 caucus of the APHA: <http://spiritof1848.org/>). Relatedly, AI4SG advocates could support enforcement of existing laws and public rights to benefit from science and technology, such as Bayh-Dole ‘march in’ rights to secure public access to publicly funded

health treatments (UCL 2018) as well as antitrust laws, alongside groups like the Economic Security Project (<https://www.economicsecurityproject.org>). They could advocate for more democratic ownership and governance of ‘sharing economy’ platforms, such as Airbnb and Uber, with organizations contributing to the platform cooperativism movement (e.g. <https://platform.coop/> and <https://sassafras.coop>). They could also advocate for innovative new institutions, such as investing profits from publicly funded technologies in public tech dividends, to support publics’ access to upstream social determinants of health. As sociologists Manuel Pastor, Chris Benner, and colleagues have written, this would be a way to ‘more directly link returns to the risk the public sector absorbs in these new innovations ... Such a ‘technology dividend’ could support a universal basic income fund, which would mitigate risk for those working through the vagaries of employment shifts engendered by innovation and technological change’ (Pastor et al. 2018, 34).

Beyond technology dividends, however, sociologist and labour scholar Annette Bernhardt underscores diverse publics’ right to shape technology development from the outset, from technologies impinging on work and labour to algorithms affecting lending, hiring and sentencing decisions (Bernhardt 2017). She outlines strategies ranging from the mitigation of technological effects, to collective bargaining, to publics having ‘a seat at the table when decisions are made over which technologies are developed in the first place, and in pursuit of which goals’, citing Germany’s collective bargaining and ‘social partners system’ in which the government ‘actively collaborat[es] with employers and labor to make its manufacturing sector a leader in technology and preserve[s] a role for workers’ (Bernhardt 2017). Such multi-stakeholder governance possibilities, reinforced by regulatory enforcement, are far removed from the USA’s current industry-led, self-regulating ‘Partnership on AI to Benefit People and Society’. They are also not among the international comparisons most frequently cited in US policy-makers’ discussions of AI. This paper just begins to scratch the surface of many of these issues. Yet ultimately, abundant data suggest that new arrangements and social transformations are urgently needed, not only for greater public benefit from and democratic governance of technologies, but for democracy – with digital and analog worlds oriented toward social justice and health equity for all.

Acknowledgments

The author is grateful to Shunryu Colin Garvey and the anonymous reviewers, who all contributed extremely helpful, generous feedback on this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Cheryl Holzmeier is a sociologist who completed her Ph.D. at UC-Berkeley. She is currently a Research Fellow affiliated with the Institute for Social Transformation at UC-Santa Cruz. Her research focuses on the intersections of science, technology, social justice, and health equity.

References

- Abernethy, J., A. Chojnacki, A. Farahi, E. Schwartz, and J. Webb. 2018. "ActiveRemediation: The Search for Lead Pipes in Flint, Michigan," ArXiv [preprint]. <https://arxiv.org/abs/1806.10692>.
- Abnoui, F., J. Rumsfeld, and H. Krumholz. 2019. "Social Determinants of Health in the Digital Age: Determining the Source Code for Nurture." *Journal of the American Medical Association* 321 (3): 247–248.
- Acemoglu, D., A. Manera, and P. Restrep. 2020. "Does the U.S. Tax Code Favor Automation?" Brookings Institution Spring Conference Draft. <https://www.brookings.edu/wp-content/uploads/2020/03/Acemoglu-et-al-Conference-Draft.pdf>.
- Achiume, E. T. 2020. "Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis," Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance. United Nations Office of the High Commissioner for Human Rights. <https://undocs.org/en/A/HRC/44/57>.
- Adams, R. 2021. "Decolonising Artificial Intelligence: A Theoretical Critique." *Interdisciplinary Science Reviews*.
- Alberti, P., A. Bonham, and D. Kirch. 2013. "Making Equity a Value in Value-Based Health Care." *Academic Medicine* 88 (11): 1619–1623.
- Albornoz, D. 2018. "Reimagining Open Science Through a Feminist Lens," *Medium*. <https://medium.com/@denalbz/reimagining-open-science-through-a-feminist-lens-546f3d10fa65>.
- American Public Health Association. "Health Equity". <https://www.apha.org/topics-and-issues/health-equity>.
- Amrute, S. 2019. "Tech Colonialism Today," EPIC2019 Keynote Address. <https://points.datasociety.net/tech-colonialism-today-9633a9cb00ad>.
- Anderson, M. K. 2005. *Tending the Wild: Native American Knowledge and the Management of California's Natural Resources*. Berkeley, CA: University of California Press.
- Asdal, K., B. Brenna, and I. Moser, eds. 2007. *Technoscience: The Politics of Interventions*. Oslo, Norway: Oslo Academic Press. Unipub Norway.
- Bassett, M. 2015. "#BlackLivesMatter — A Challenge to the Medical and Public Health Communities." *New England Journal of Medicine* 372 (12): 1085–1087.
- Bayer, R., and S. Galea. 2015. "Public Health in the Precision-Medicine Era." *The New England Journal of Medicine* 373 (6): 499–501.
- Beckfield, J. 2018. *Political Sociology and the People's Health*. New York, NY: Oxford University Press.
- Benjamin, R. 2019a. *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity Press.
- Benjamin, R. 2019b. "Assessing Risk, Automating Racism." *Science* 366 (6464): 421–422.
- Bernhardt, A. 2017. "Expanding the Goal of Innovation," in Forum on "Basic Income in a Just Society," *Boston Review*, Spring 2017. <http://bostonreview.net/forum/basic-income-just-society/annette-bernhardt-expanding-goal-innovation>.

- Birn, A. 2014. "Philanthrocapitalism, Past and Present: The Rockefeller Foundation, the Gates Foundation, and the Setting(s) of the International/Global Health Agenda." *Hypothesis* 12 (1): 1–27.
- boyd, d., and K. Crawford. 2012. "CRITICAL QUESTIONS FOR BIG DATA: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–679.
- Boyd, R., E. Lindo, L. Weeks, and M. McLemore. 2020. "On Racism: A New Standard for Publishing On Racial Health Inequities," *Health Affairs Blog*. <https://www.healthaffairs.org/doi/10.1377/hblog20200630.939347/full/>.
- Bozeman, B. 2020. "Public Value Science." *Issues in Science and Technology* 36 (4): 34–41.
- Bradley, E., and L. Taylor. 2013. *The American Healthcare Paradox: Why Spending More is Getting Us Less*. New York: Public Affairs Books.
- Brandt, A., and M. Gardner. 2000. "Antagonism and Accommodation: Interpreting the Relationship Between Public Health and Medicine in the United States During the 20th Century." *American Journal of Public Health* 90 (5): 707–715.
- Braveman, P., S. Egerter, S. Woolf, and J. Marks. 2011. "When Do We Know Enough to Recommend Action on the Social Determinants of Health?" *American Journal of Preventive Medicine* 40 (1, Suppl. 1): S58–66.
- Braveman, P., and L. Gottlieb. 2014. "The Social Determinants of Health: It's Time to Consider the Causes of the Causes." *Public Health Reports* 129 (Suppl 2): 19–31.
- Brown, P. 1992. "Popular Epidemiology and Toxic Waste Contamination: Lay and Professional Ways of Knowing." *Journal of Health and Social Behavior* 33 (3): 267–81.
- Brown, P., R. Morello-Frosch, and S. Zavestoski. 2011. *Contested Illnesses: Citizens, Science, and Health Social Movements*. Berkeley, CA: University of California Press.
- Burton, S. D., T. Mahfoud, C. Aicardi, and N. Rose. 2021. "Clinical Translation of Computational Brain Models: Understanding the Salience of Trust in Clinician-Researcher Relationships." *Interdisciplinary Science Reviews*.
- California Department of Public Health (CDPH), Office of Health Equity. 2015. "Portrait of Promise: The California Statewide Plan to Promote Health and Mental Health Equity." Sacramento, CA: Office of Health Equity, CDPH. [https://www.cdph.ca.gov/Programs/OHE/CDPH%20Document%20Library/Accessible-CDPH_OHE_Disparity_Report_Final%20\(2\).pdf](https://www.cdph.ca.gov/Programs/OHE/CDPH%20Document%20Library/Accessible-CDPH_OHE_Disparity_Report_Final%20(2).pdf).
- California Precision Medicine Advisory Committee. 2019. "Precision Medicine: An Action Plan for California". https://opr.ca.gov/docs/20190107-Precision_Medicine_An_Action_Plan_for_California.pdf.
- Cath, C., S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi. 2017. "Artificial Intelligence and the 'Good Society': the US, EU, and UK Approach." *Science and Engineering Ethics* 24 (2): 505–528.
- Chadha, N., B. Lim, M. Kane, and B. Rowland. 2020. *Toward the Abolition of Biological Race in Medicine: Transforming Clinical Education, Research, and Practice*, Othering and Belonging Institute at UC-Berkeley and Institute for Healing and Justice in Medicine. <https://www.instituteforhealingandjustice.org/download-the-report-here>.
- Chen, I., P. Szolovits, and M. Ghassemi. 2019. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?" *AMA Journal of Ethics* 21 (2): E167–179.
- Chowkwanyun, M., R. Bayer, and S. Galea. 2018. "'Precision' Public Health – Between Novelty and Hype." *The New England Journal of Medicine* 379 (15): 1398–1400.
- Chowkwanyun, M., R. Bayer, and S. Galea. 2019. "Precision Public Health: Pitfalls and Promises." *The Lancet* 393 (10183): 1801.
- Chui, M., M. Harryson, J. Manyika, R. Roberts, R. Chung, A. van Heteren, and P. Nel. 2018. "Notes from the AI Frontier: Applying AI for Social Good," McKinsey Global Institute,

- McKinsey & Company. <https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good>.
- Coburn, D. 2010. "Beyond the Income Inequality Hypothesis: Class, Neo-Liberalism, and Health Inequalities." In *Tackling Health Inequities Through Public Health Practice: Theory to Action, 2nd Edition; A Project of the National Association of City and County Health Officials*, edited by R. Hofrichter, and R. Bhatia, 196–224. New York: Oxford University Press.
- Cognizant. 2019. "The Social Determinants of Health: Applying AI & Machine Learning to Achieve Whole Person Care". <https://www.cognizant.com/whitepapers/the-social-determinants-of-health-applying-ai-and-machine-learning-to-achieve-whole-person-care-codex4379.pdf>.
- Comstock, J. 2018. "SkinVision gets \$7.6M to continue expanding skin cancer app," *Mobile Health News*. <https://www.mobihealthnews.com/content/skinvision-gets-76m-continue-expanding-skin-cancer-app>.
- Cooper, R., and N. Paneth. 2020a. "Will Precision Medicine Lead to a Healthier Population?" *Issues in Science and Technology* 36 (2): 64–71.
- Cooper, R., and N. Paneth. 2020b. "Precision medicine: course correction urgently needed," *STAT News*, 3 March 2020. <https://www.statnews.com/2020/03/03/precision-medicine-course-correction-urgently-needed/>.
- Corburn, J. 2005. *Street Science: Community Knowledge and Environmental Health Justice*. Cambridge, MA: The MIT Press.
- DeVoe, J., A. Geller, and Y. Negussie. 2019. *Vibrant and Healthy Kids: Aligning Science, Practice, and Policy to Advance Health Equity*. Washington, D.C.: National Academy Press.
- Engstrom, D., D. Ho, C. Sharkey, and M. Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, a report submitted to the Administrative Conference of the United States. <https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf>.
- Epstein, S. 1995. "The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials." *Science, Technology, & Human Values* 20 (4): 408–437.
- Espeland, W. 1998. *The Struggle for Water: Politics, Rationality, and Identity in the American Southwest*. Chicago, IL: University of Chicago Press.
- Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.
- Fairchild, A., D. Rosner, J. Colgrove, R. Bayer, and L. Fried. 2010. "The EXODUS of Public Health: What History Can Tell Us About the Future." *American Journal of Public Health* 100 (1): 54–63.
- Ferguson, J. 1994. *The Anti-Politics Machine: 'Development' and Bureaucratic Power in Lesotho*. Minneapolis, MN: University of Minnesota Press.
- Ferryman, K., and M. Pitcan. 2018. *Fairness in Precision Medicine*. New York: Data & Society. <https://datasociety.net/library/fairness-in-precision-medicine/>.
- Ferryman, K., and R. Winn. 2018. "Artificial Intelligence Can Entrench Disparities – Here's What We Must Do," *The Cancer Letter*. https://cancerletter.com/articles/20181116_1/.
- Figueroa, J., A. Frakt, and A. Jha. 2020. "Addressing Social Determinants of Health: Time for a Polysocial Risk Score." *Journal of the American Medical Association* 323 (16): 1553–1554.
- Floridi, L., J. Cows, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, et al. 2018. "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28 (4): 689–707.

- Floridi, L., J. Cows, T. King, and M. Taddeo. 2020. "How to Design AI for Social Good: Seven Essential Factors." *Science and Engineering Ethics* 26: 1771–1796.
- Fortun, M. 2008. *Promising Genomics: Iceland and deCODE Genetics in a World of Speculation*. Berkeley, CA: University of California Press.
- Funk, J. 2019. "What's Behind Technological Hype?" *Issues in Science and Technology* 36 (1): 36–42.
- Galea, S. 2019. *Well: What We Need to Talk About When We Talk About Health*. Oxford, UK: Oxford University Press.
- Garvey, C. 2018. "Interview with Colin Garvey, Rensselaer Polytechnic Institute. *Artificial Intelligence and Systems Medicine Convergence*." *OMICS: A Journal of Integrative Biology* 22 (2): 130–132.
- Google AI Impact Challenge. 2019. "Accelerating social good with artificial intelligence: Insights from the Google AI Impact Challenge". https://services.google.com/fh/files/misc/accelerating_social_good_with_artificial_intelligence_google_ai_impact_challenge.pdf.
- Gray, M., and S. Suri. 2019. *Ghostwork: How to Stop Silicon Valley from Building a New Global Underclass*. Boston, MA: Houghton Mifflin Harcourt.
- Hager, G., A. Drobni, F. Fang, R. Ghani, A. Greenwald, T. Lyons, D. Parkes, et al. 2017. *Artificial Intelligence for Social Good*. Washington, DC: Computing Community Consortium. <https://arxiv.org/ftp/arxiv/papers/1901/1901.05406.pdf>.
- Hanemaayer, A. 2021. "Don't Touch My Stuff: Historicizing Resistance to AI and Algorithmic Computer Technology in Medicine." *Interdisciplinary Science Reviews*.
- Harvard Heart Letter. 2019. "Smartphone Apps for Managing Heart Disease". <https://www.health.harvard.edu/heart-health/smartphone-apps-for-managing-heart-disease>.
- Hatton, E. 2020. *Coerced: Work Under Threat of Punishment*. Berkeley, CA: University of California Press.
- Hawn Nelson, A., D. Jenkins, S. Zanti, M. Katz, E. Berkowitz, T. Burnett, D. Culhane 2020. *A Toolkit for Centering Racial Equity Throughout Data Integration*, Actionable Intelligence for Social Policy, University of Pennsylvania. <https://www.aisp.upenn.edu/equity-toolkit/>.
- Heaven, W. 2020. "Predictive policing algorithms are racist. They need to be dismantled," *MIT Technology Review*. <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>.
- Ho, D., S. Quake, E. McCabe, W. J. Chng, E. Chow, X. Ding, B. Gelb, et al. 2020. "Enabling Technologies for Personalized and Precision Medicine." *Trends in Biotechnology* 38 (5): 497–518.
- Hofrichter, R., and R. Bhatia. 2010. *Tackling Health Inequities Through Public Health Practice: Theory to Action, 2nd Edition; A Project of the National Association of City and County Health Officials*. New York: Oxford University Press.
- Holzmeyer, C. 2017. "Investigating STEM – Health Equity as Touchstone for the Future," *Boom California*. <https://boomcalifornia.com/2017/09/06/investigating-stem-health-equity-as-touchstone-for-the-future/>.
- Holzmeyer, C. 2018a. "Open Science Initiatives: Challenges for Public Health Promotion." *Health Promotion International* 34 (3): 624–633.
- Holzmeyer, C. 2018b. "The Crockett-Rodeo Real-Time Health Monitoring Study: Summary of Findings," 27 June 2018. https://drive.google.com/file/d/1QCCZlpRVDbCcSdpA__3LMb1Z_M9SjIY8/view.
- Holzmeyer, C. 2018c. "Wider Worlds of Research for Health Equity: Public Health NGOs as Stakeholders in Open Access Ecosystems." *The Journal of Community Informatics* 14 (2–3): 1–16.

- Huesemann, M., and J. Huesemann. 2011. *Techno-Fix: Why Technology Won't Save Us or the Environment*. BC, Canada: New Society Publishers.
- Interlandi, J. 2020. "The U.S. Approach to Public Health: Neglect, Panic, Repeat," *The New York Times*, 9 April 2020. <https://www.nytimes.com/2020/04/09/opinion/sunday/coronavirus-public-health-system-us.html>.
- Iton, A. 2010. "The Ethics of the Medical Model in Addressing the Root Causes of Health Disparities in Local Public Health Practice." In *Tackling Health Inequities Through Public Health Practice: Theory to Action, 2nd Edition; A Project of the National Association of City and County Health Officials*, edited by R. Hofrichter, and R. Bhatia, 509–516. New York: Oxford University Press.
- Iton, A. 2017. "Rebuilding Our Social Compact," Building Healthy Communities Blog, The California Endowment. <https://www.calendow.org/rebuilding-social-compact/>.
- Iton, A. 2020. "'Policy violence' against people of color increases COVID-19's deadly toll," *The Sacramento Bee*. <https://www.sacbee.com/opinion/california-forum/article242010461.html>.
- Jobin, A., M. Ienca, and E. Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (2019): 389–399.
- Khoury, M., and J. Ioannidis. 2014. "Big Data Meets Public Health." *Science* 346 (6213): 1054–1055.
- Krieger, N. 2011. *Epidemiology and the People's Health: Theory and Context*. New York: Oxford University Press.
- Krieger, N. 2020. "ENOUGH: COVID-19, Structural Racism, Police Brutality, Plutocracy, Climate Change—and Time for Health Justice, Democratic Governance, and an Equitable, Sustainable Future." *American Journal of Public Health* 110 (11): 1620–1623.
- Krieger, N., and E. Birn. 1998. "A Vision of Social Justice as the Foundation of Public Health: Commemorating 150 Years of the Spirit of 1848." *American Journal of Public Health* 88 (11): 1603–1606.
- Krieger, N., and G. D. Smith. 2016. "The Tale Wagged by the DAG: Broadening the Scope of Causal Inference and Explanation for Epidemiology." *International Journal of Epidemiology* 45 (6): 1787–1808.
- Kristof, N. 2020. "'We're No. 28! And Dropping!'" *The New York Times*, 9 September 2020. <https://www.nytimes.com/2020/09/09/opinion/united-states-social-progress.html>.
- Kuch, D., M. Kearnes, and K. Gulson. 2020. "The Promise of Precision: Datafication in Medicine, Agriculture and Education." *Policy Studies* 41 (5): 527–546.
- Latonero, M. 2019. "AI For Good Is Often Bad," *Wired*. <https://www.wired.com/story/opinion-ai-for-good-is-often-bad/>.
- Lazonick, W., and M. Mazzucato. 2013. "The Risk-Reward Nexus in the Innovation-Inequality Relationship: Who Takes the Risks? Who Gets the Rewards?." *Industrial and Corporate Change* 22 (4): 1093–1128.
- Ledford, H. 2020. "Computing Humanity: How Data from Facebook, Twitter and Other Sources are Revolutionizing Social Science." *Nature* 582: 328–330.
- Lewis, T., S. P. Gangadharan, M. Saba, and T. Petty. 2018. *Digital Defense Playbook: Community Power Tools for Reclaiming Data*. Detroit, MI: Our Data Bodies Project. <https://www.odbproject.org/tools/>.
- Lélé, S., and R. Norgaard. 2005. "Practicing Interdisciplinarity." *BioScience* 55 (11): 967–975.
- Lou, Z., A. Raval, M. Young, and S. Appel. 2020. *Resilience Before Disaster: The Need to Build Equitable, Community-Driven Social Infrastructure*, Asian Pacific Environmental Network (APEN), Service Employees International Union (SEIU) California, and BlueGreen Alliance. <https://apen4ej.org/resilience-before-disaster/>.

- Lown Institute. 2019. *California's Health Care Paradox: Too much health care spending may lead to poor community health*. <https://lowninstitute.org/reports/californias-health-care-paradox-2/>.
- Maani, N., and S. Galea. 2020a. "The Role of Physicians in Addressing Social Determinants of Health." *Journal of the American Medical Association* 323 (16): 1551–1552.
- Maani, N., and S. Galea. 2020b. "COVID-19 and Underinvestment in the Health of the U.S. Population." *Milbank Quarterly* 98 (2): 239–249.
- Martin, B. 2006. "Strategies for Alternative Science." In *The New Political Sociology of Science: Institutions, Networks, and Power*, edited by S. Frickel, and K. Moore, 272–298. Madison, WI: University of Wisconsin Press.
- Mazzucato, M. 2013. *The Entrepreneurial State: Debunking Public vs. Private Sector Myths*. London, UK: Anthem Press.
- Mazzucato, M. 2018. *The Value of Everything: Making and Taking in the Global Economy*. UK: Allen Lane.
- Metzl, J., and H. Hansen. 2014. "Structural Competency: Theorizing a New Medical Engagement With Stigma and Inequality." *Social Science & Medicine* 103 (1): 126–133.
- Milner, Y. 2018. "An Open Letter to Facebook from the Data for Black Lives Movement," *Medium*. <https://medium.com/@YESHICAN/an-open-letter-to-facebook-from-the-data-for-black-lives-movement-81e693c6b46c>.
- Milner, Y. 2020. "We Will Not Allow the Weaponization of COVID-19 Data," *Medium*. <https://medium.com/@YESHICAN/we-will-not-allow-the-weaponization-of-covid-19-data-e775d31991c>.
- Mittelstadt, B., and L. Floridi. 2016. "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts." *Science and Engineering Ethics* 22 (2): 303–341.
- Mohamed, S., M. Png, and W. Isaac. 2020. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *Philosophy & Technology* 405: 1–28.
- Moore, J. 2019. "AI for Not Bad." *Frontiers in Big Data* 2 (32): 1–7.
- Movement for Black Lives. 2020. "Vision for Black Lives: 2020 Policy Platform". <https://m4bl.org/policy-platforms/>.
- National Domestic Workers Alliance. 2020. "Summary of the National Domestic Workers Bill of Rights". <https://www.domesticworkers.org/sites/all/themes/NDWA2017/images/LearnMore.pdf>.
- Neff, J. 2020. "Structural Competency for Healthcare Providers," unpublished manuscript, shared with the Structural Competency Working Group (<https://www.structcomp.org/>).
- Noble, S. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- Nuffield Council on Bioethics. 2018. "Bioethics Briefing Note: Artificial intelligence (AI) in healthcare and research". <https://www.nuffieldbioethics.org/publications/ai-in-healthcare-and-research>.
- Obasogie, O., and M. Darnovsky. 2018. *Beyond Bioethics: Toward a New Biopolitics*. Berkeley, CA: University of California Press.
- Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Patients." *Science* 366 (6464): 447–453.
- Ojo, A., P. Erfani, and N. Shah. 2020. "Value-Based Health Care Must Value Black Lives," *Health Affairs Blog*. <https://www.healthaffairs.org/doi/10.1377/hblog20200831.419320/full/>.
- Onuoha, M. and Mother Cyborg (Diana Nucera). 2018. *A People's Guide to AI*. Detroit, MI: Allied Media Projects. <https://www.alliedmedia.org/peoples-ai>.
- Osther, K. 2019. "You Don't Want Facebook Involved With Your Health Care." *Slate*. <https://slate.com/technology/2019/09/social-determinants-health-facebook-google.html>.

- Panch, T., J. Pearson-Stuttard, F. Greaves, and R. Atun. 2019. "Artificial Intelligence: Opportunities and Risks for Public Health." *The Lancet* 1 (1): e13–e14.
- Pastor, M., C. Benner, A. Sharma, E. Muña, S. Rosenquist, and V. Carter. 2018. *From Resistance to Renewal: A 12-Step Program for Innovation and Inclusion in the California Economy*, Program for Environmental and Regional Equity, University of Southern California and Institute for Social Transformation, UC-Santa Cruz. <https://dornsife.usc.edu/pere/california-economy>.
- Petty, T., M. Saba, T. Lewis, S. P. Gangadharan, and V. Eubanks. 2018. "Our Data Bodies: Reclaiming Our Data (Interim Report)." Detroit, MI: Our Data Bodies Project. <https://www.odbproject.org/tools/>.
- Poor People's Campaign. 2020. "A Moral Policy Agenda to Heal and Transform America: The Poor People's Jubilee Platform". <https://www.poorpeoplescampaign.org/about/jubilee-platform/>.
- Pūras, D. 2020. "Right of Everyone to the Enjoyment of the Highest Attainable Standard of Physical and Mental Health," Report of the Special Rapporteur on the right of everyone to the enjoyment of the highest attainable standard of physical and mental health. United Nations Office of the High Commissioner for Human Rights. <https://undocs.org/en/A/HRC/44/48>.
- Rolnick, D., P. Donti, L. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. Ross et al. 2019. "Tackling Climate Change with Machine Learning," *ArXiv [preprint]*. <https://arxiv.org/abs/1906.05433>.
- Rudolph, L., J. Caplan, K. Ben-Moshe, and L. Dillon. 2013. *Health in All Policies: A Guide for State and Local Governments*. Washington, DC: American Public Health Association and Public Health Institute.
- Sachs, J., J. Stiglitz, M. Mazzucato, C. Brown, I. Dutta-Gupta, R. Reich, G. Zucman, et al. 2020. "Letter from economists: to rebuild our world, we must end the carbon economy," *The Guardian*. <https://www.theguardian.com/commentisfree/2020/aug/04/economists-letter-carbon-economy-climate-change-rebuild>.
- Sarewitz, D. 2004. "How Science Makes Environmental Controversies Worse." *Environmental Science & Policy* 7 (5): 385–403.
- Sarewitz, D. 2015. "CRISPR – Science Can't Solve It." *Nature* 522 (7557): 413–414.
- Sarewitz, D., and R. Nelson. 2008. "Three Rules for Technological Fixes." *Nature* 456 (7224): 871–872.
- Schork, N. 2019. "Artificial Intelligence and Personalized Medicine." In *Precision Medicine in Cancer Therapy*, edited by D. Von Hoff, and H. Han, 265–283. Berlin, Germany: Springer.
- Serkez, Y. 2020. "Who Is Most Likely to Die From the Coronavirus?" *The New York Times*, 4 June 2020. <https://www.nytimes.com/interactive/2020/06/04/opinion/coronavirus-health-race-inequality.html>.
- Shi, Z. R., C. Wang, and F. Fang. 2020. "Artificial Intelligence for Social Good: A Survey," *ArXiv [preprint]*. <https://arxiv.org/abs/2001.01818>.
- Shonkoff, J., and D. Phillips. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development*. Washington, D.C: National Academy Press.
- Silverstein, M., H. Hsu, and A. Bell. 2019. "Addressing Social Determinants to Improve Population Health: The Balance Between Clinical Care and Public Health." *Journal of the American Medical Association* 322 (24): 2379–2380.
- Singer, N. 2017. "How Big Tech is Going After Your Health Care," *The New York Times*, 26 December 2017. <https://www.nytimes.com/2017/12/26/technology/big-tech-health-care.html>.

- Singer, N. 2019. "When Apps Get Your Medical Data, Your Privacy May Go With It," *The New York Times*, 3 September 2019. <https://www.nytimes.com/2019/09/03/technology/smartphone-medical-records.html>.
- Stewart, J. 2017. "Smartphone App Shown to Accurately Diagnose Respiratory Illnesses Like COPD," *COPD News Today*. <https://copdnewstoday.com/2017/12/21/resapp-smartphone-app-accurately-diagnoses-copd-other-respiratory-illnesses/>.
- Strober, M. 2010. *Interdisciplinary Conversations: Challenging Habits of Thought*. Stanford, CA: Stanford University Press.
- Strubell, E., A. Ganesh, and A. McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP," *ArXiv [preprint]*. <https://arxiv.org/abs/1906.02243>.
- Sun, W., J. Lee, S. Zhang, C. Benyshek, M. Dokmeci, and A. Khademhosseini. 2018. "Engineering Precision Medicine," *Advanced Science* 6 (1): 1–19.
- Sunrise Movement. 2020. "Green New Deal"; <https://www.sunrisemovement.org/green-new-deal>.
- Tekisalp, L. 2020. "Beyond Hype and Innovation: AI for Social Good," Partnership on AI Blog. <https://www.partnershiponai.org/beyond-hype-and-innovation-ai-for-social-good/>.
- Tomašev, N., J. Cornebise, F. Hutter, S. Mohamed, A. Picciariello, B. Connelly, D. Belgrave, et al. 2020. "AI for Social Good: Unlocking the Opportunity for Positive Impact." *Nature Communications* 11 (2468): 1–6.
- Ulnicane, I., D. O. Eke, W. Knight, G. Ogoh, and B. C. Stahl. 2021. "Good governance as a Response to Discontents? Déjà vu, or Lessons for AI from other Emerging Technologies." *Interdisciplinary Science Reviews*.
- University College London (UCL), Institute for Innovation and Public Purpose. 2018. "The people's prescription: Re-imagining health innovation to deliver public value," IIPP Policy Report, 2018-10. London, UK: IIPP, Global Justice Now, Just Treatment, STOPAIDS. <https://www.ucl.ac.uk/bartlett/public-purpose/wp2018-10>.
- Wallack, L., and K. Thornburg. 2016. "Developmental Origins, Epigenetics, and Equity: Moving Upstream." *Maternal and Child Health Journal* 20 (5): 935–940.
- Wilkinson, R., and K. Pickett. 2019. *The Inner Level: How More Equal Societies Reduce Stress, Restore Sanity and Improve Everyone's Well-Being*. UK: Penguin Random House.
- Woodhouse, E., and D. Sarewitz. 2007. "Science Policies for Reducing Societal Inequities." *Science and Public Policy* 34 (2): 139–150.
- Wolf, S., and L. Aron. 2013. *U.S. Health in International Perspective: Shorter Lives, Poorer Health*. Washington, D.C: National Academy Press.
- Wolf, S., R. Johnson, G. Fryer, Jr., G. Rust, and D. Satcher. 2008. "The Health Impact of Resolving Racial Disparities: An Analysis of U.S. Mortality Data." *American Journal of Public Health* 98 (Suppl 1): S26–S28.
- World Health Organization. 2008. *Closing the Gap in a Generation: Health Equity Through Action on the Social Determinants of Health*; *Final Report of the Commission on Social Determinants of Health*. Geneva: World Health Organization.
- World Health Organization. 2017. "Human Rights and Health," WHO Fact Sheet. <https://www.who.int/news-room/fact-sheets/detail/human-rights-and-health>.
- Wu, H., H. Yin, H. Chen, M. Sun, X. Liu, Y. Yu, Y. Tang, et al. 2020. "A Deep Learning, Image Based Approach for Automated Diagnosis for Inflammatory Skin Diseases." *Annals of Translational Medicine* 8 (9): 581.
- Wylie, S., N. Shapiro, and M. Liboiron. 2017. "Making and Doing Politics Through Grassroots Scientific Research on the Energy and Petrochemical Industries." *Engaging Science, Technology, and Society* 3: 393–425.



Don't touch my stuff: historicising resistance to AI and algorithmic computer technologies in medicine

Ariane Hanemaayer 

Department of Sociology, Brandon University

ABSTRACT

This paper historicises the criticisms and backlash from within medicine against new computer technologies in the clinic. Physicians' reactions to proposals for the implementation of algorithmic technologies in the clinic ranged from enthusiastic to cautionary to critical from as early as the 1960s. Clinicians were suspicious of these technologies as they threatened their professional expertise. I argue that these discontent reactions from doctors demonstrate an implicit struggle for authority over clinical spaces and with regards to medicine's place within society more generally. Drawing on Foucault's concept of discursive rules and their function within a closed community, I recover the forgotten debate to include or reject AI and its predecessor technologies of expert systems and neural networks. This paper explains how and why justifications for and against the applicability of AI to the clinic are underpinned by questions of medical authority. I conclude with an inquiry into the transformative possibilities of partisanship.

KEYWORDS

AI; medicine; Foucault; clinical authority; decision-making; counter-human science; clinical expertise; rarefaction

Research and funding for the use of computers in biomedical and laboratory research began in the 1950s. By the late 1960s, the computer had been proposed as a helpful system for clinical decision-making, coming far afield of research (e.g. Taylor 1967). From that time on, a nerve was struck: The role of emerging technologies in decision making – if any – became a contentious topic among clinicians. By the 1990s, special forums in medicine's flagship journals were dedicated to the debate. Consider a few statements of clinicians' discontent:

- Computers lack power of natural inference and an illative sense; so although they have reliable memories and may be of great help in diagnosis and management, they can never substitute for the able physician. (Beautyman 1982, 932)
- The very name [artificial intelligence] suggests kinship with the deepest secrets of human thought and ... this illusion is one which the purveyors

CONTACT Ariane Hanemaayer  hanemaayera@brandonu.ca

© 2020 Institute of Materials, Minerals and Mining Published by Taylor & Francis on behalf of the Institute

of this new product are not averse to propagating ... The truth is that, apart from a passing geometrical similarity, there is no evidence that ANNs bear any relation to the workings of human minds. (Signorini and Slattery 1995, 1500)

- No statistical rules, numbers or algorithms can replace clinical judgment in the interpretation of the results of clinical research. (Ioannidis and Lau 1996, 756)
- Clinical certainty promised by highly sophisticated technology and evidence-based medical algorithms is an illusion that cannot dispense with personal experience. (Benítez-Bribiesca 2000, 761)
- Computers aren't able to care for patients in the sense of showing devotion or concern for the other as a person, because they are not people and do not care about anything. (Goldhahn, Rampton, and Spinus 2018)

What kind of reasoning unifies these statements? And what is at stake in the struggle over the role of computer technology in the clinic?

This article historicises resistance to new computer technologies in the clinic from within medicine. Physicians' reactions to proposals for the clinical implementation of AI ranged from enthusiastic to cautionary to critical. Its acceptance, however, has been slowed for a number of physician-identified reasons: inadequate computer technology; insufficient data for 'training' artificial neural networks; clinical decisions are not easily rule-guided by algorithms; and questions about legal and medical responsibilities in using 'black box' AI technologies remain unanswered. Clinicians were and continue to be suspicious of AI, for it threatens their professional expertise. Despite the range of reasons provided by physicians for their reticence, I argue that doctors' discontent demonstrates the existence of an implicit struggle for authority over clinical truth – and the place of medicine within society.

This article begins with an explanation of the theoretical frame I use to analyse physicians' responses to computer technology in the clinic. I conceptualize doctors' discontent as an outcome of *rarefaction*, the process of making the authority of clinical truth rare or less accessible to those outside the medical community. The body of the article analyses statements from the archive with respect to the four elements of rarefaction: *ritual*, *closed community*, *doctrine*, and *social conflict*. These elements allow us to understand discontent as a struggle for authority in the clinic. After discussing some implications of my analysis, I conclude by suggesting the basis for partisanship among clinical actors, in lieu of discontent.

Framing statements of discontent

Although twenty-first century medicine has changed significantly from the way it was practised in the 1960s, the 'play of dominations' within medical discourse

(Walters 2012, 133; c.f. Foucault 1998, 376) is notable for the strong statements of discontent that appear alongside the usual embrace of computer technologies, such as AI and machine learning (ML). One physician, for example, recently argued for the importance of human touch in the clinic (Horton 2019). While resistance to AI does not define the contemporary clinic, doubts persist. Recent efforts affirm the need for further consideration of the disconnect between health systems, regulation, and technological innovation (Kickbusch et al. 2019). That calls to (re)define and regulate 'digital health' continue to be made even today (Bayram et al. 2020) signal an ongoing story of medical resistance, a forgotten struggle that remains to be historicised. This paper makes that intervention.

The struggles to resist new technologies require recovery. Doing so provides our historical record with the discomfort of clinical discontents who sought to have a say about computer technologies, their role in society and place in the clinic. Starting from the premise that history is the outcome of struggle, I begin by examining the principles that unify the clinically discontented. Drawing on Foucault's concept of discursive rules and their function within a closed community, I frame my analysis of the debate to include or reject AI and its predecessor technologies of expert systems and neural networks as an attempt to control the authorship of truth in the clinic. I consider discourse that constitutes this debate to be a collection of statements, their production 'at once controlled, selected, organized and redistributed according to a certain number of procedures, whose role is to avert its powers and its dangers, to cope with chance events, to evade its ponderous, awesome materiality' (Foucault 1972a, 216). These principles function to separate the true from the false. Further, discourses prescribe how truth can be used and how it is useful. Statements in medical journals by clinical physicians, medical technicians and scientists recommend how and by whom truth is made in medicine, as well as who and what are able to produce diagnoses, prognoses and predictions.

My approach is rooted in the critical tradition that 'analyses the processes of rarefaction, consolidation and unification in discourse' (Foucault 1972, 233). Rarefaction consists of four processes: ritual, closed community, doctrine and social conflict. My analysis of the medical literature is organized accordingly. The method for the collection of statements consisted of conducting historical research. PubMed and the Wellcome Trust database of medical journals were searched both online and, in the latter case, onsite. Searches were conducted using the following terms: 'computer technologies', 'AI' and 'expert systems'. Articles were further sorted based on their relevance to the research question about the usefulness, acceptance or rejection of computer technologies. Paper copies of undigitized articles were retrieved at the Wellcome Trust Library, and images taken. This process yielded a sample of 99 peer reviewed journal

articles and an additional 613 pages of copied materials from original research, conference proceedings and opinion pieces.

I define statements as events within a discourse (Hanemaayer 2019a, 225; see also Foucault 1972b, 4). The archive contains a set of many discourses ‘actually pronounced’, which have a function in the institutions where they are stated, and often change over time (Hanemaayer 2019a, 225; Foucault 1972b, 57). Documents within the archive contain statements; in this case, statements enunciated by doctors, clinicians, medical researchers and medical systems experts. In my project, those statements that dealt with the kinds of discursive and institutional support lent to the implementation of and investment in computer technologies were selected. This collection of documents formed the sample of my analysis.

Given that my project takes the archive as a nominal index for the rationality that organizes human activity, I examine the statements of medical discourse as indicative of a struggle to constitute what does and does not belong in the clinic and who does and does not have access to clinical knowledge. This conceptualization follows Foucault’s definition of the rarefaction of discourse, defined as the conditions under which someone can employ it, the rules imposed on those who deploy it, and the exclusive access to truth (1972a, 224). My paper uses this framework, first, to explain via the elements of the rarefaction of discourse why clinical statements refuse to share their discursive access to truth with machine processing, and, second, to briefly discuss the disruption AI technology poses for medical discourses by describing what is at stake in the functions of control, internal rules, the author and the discipline.

Rarefaction in medical decision-making

Physicians’ statements about decision making and the nature of clinical work are organized by the process of rarefaction. Following Osborne, I analyse ‘the disposal of facts in a particular way so as to produce a particular picture of things’ (1999, 58). I show how physicians, as authors, use clinical reason to relate the future of ML in a way that excludes its authority or reliability in decision making. Physicians select facts about the nature of clinical work in order to raise questions about the authority to speak truth. While the reasons and facts vary in content, these principles reveal the underlying struggle to rarefy medical truth.

Ritual

The first element of rarefaction concerns the significance of the words used: ‘Ritual defines the qualifications required of the speaker ... it lays down the gestures to be made, behaviour, circumstances, and the whole range of signs that must accompany discourse’ (Foucault 1972a, 225). With respect to the clinic, this encompasses the clinicians’ responsibility to make a good decision, as

well as to the rapport between patient and physician. The following three examples illustrate the role of ritual in securing the authority of clinical decision making for the physician.

In the 1980s, the discourse contained statements extolling the benefits of computer technology that could enable information recall for physicians. Holding all the information that could possibly be related to any single decision in one's mind was regarded to be difficult if not impossible. Despite the benefits of computerized information retrieval, however, knowing how to assemble this information was described as a feat only a physician could accomplish:

The experienced clinician knows the value of the history, examination, and investigations in solving the problem and will pick his [*sic*] way through them in a manner which will not be tethered to a branching tree. After finding certain signs, he may go backwards and delve further into the history, even after completing the clinical part of the consultation and learning the results of certain investigations he may go back yet further. Classically, an algorithm does not do this; it is non-commutative – it does not permit changes in the order in which bits of information are sought because solution of the problem depends upon the order in which they are found. (Campbell 1987, 850)

In this quotation, a clinician explains that diagnosis is a process containing a variety of behaviours, such as taking a history more than once during the course of consultation, that are associated with good decision making. Further, the 'non-communicative' elements that physicians can recognize and machines could not (at the time of his writing) show that ritual is at work, as speaking and problem solving rely as much on what the physician does and how they think as much as the information or words themselves.

In addition to the role of information and how one acts with regard to its collection and recollection is the importance of responsibility in decision making.

The clinician has to consider not only the differential diagnosis but also the potential risks, benefits, and costs of following a particular management strategy. To achieve this aim the inexperienced clinician needs the sense of perspective that a senior clinician has acquired through experience. Merely feeding clinical data into a computer and reading the result, irrespective of the method used to derive it, could undermine the clinician's ability to take personal responsibility for clinical decisions. (Dodds 1995, 1500)

This physician's justification for the rarefaction of discourse in the clinic is the qualification of the speaker. The experience of the clinician and the responsibility associated with the outcomes of the decision are to be found here. While at the time of writing algorithms may not have been as advanced in terms of assessing risk, outcome, cost and so on, these concerns continue to be expressed in the discourse. The importance of the non-spoken and the ability to weigh the divergent aspects of diagnosis in the clinic is put into the following words:

Technical knowledge cannot entirely describe the sickness situation of any single patient. A deliberative patient-physician relationship characterised by associative and lateral thinking is important for healing, particularly for complex conditions and when there is a high risk of adverse effects, because individual patients' preferences differ. There are no algorithms for such situations, which change depending on emotions, non-verbal communication, values, personal preferences, prevailing social circumstances, and so on. (Goldhahn, Rampton, and Spinus 2018, k4563)

These authors defend the authority of the physician by making reference to the broader social circumstances of the clinic, which is seen to belong to medical practitioners, not algorithms. Here risk is situated in relation to the whole patient, their preferences, emotions, and other aspects of the clinical encounter which are not readily coded into information for a machine.

Closed community

The second element of rarefaction concerns the boundaries of the clinic and those whom belong within them. Closed community refers to the circulation of discourse: who can express it, distribute it, and preserve or reproduce it (Foucault 1972a, 225). By the 1990s, the importation of computer science from the laboratory to the clinic called into question who belonged in the latter. This was expressed by one physician who explained the challenge of keeping up with two fields at the same time: 'cutting edge research in biology/biotechnology is seriously challenging even the best people in medical informatics' (Lindberg 1990, 10). Coding machines to use clinical and biomedical data to make decisions brought new 'speakers' into the clinic. Those who knew the machines and the science threatened to displace the clinician who may not be keeping up.

A decade later, some physicians accepted new technologies, albeit with boundaries and hierarchies of authority clearly defined. For example, clinical trials tested the effectiveness of new computer technologies as decision aids in the clinic, such as ISABEL, a clinical decision-making support system for pediatric care introduced in the early 2000s.

Careful consideration needs to be given as to whether results of randomised trials alone will convincingly show whether diagnostic support systems are of value in all clinical settings, especially in emergency situations. A prerequisite for the successful functioning of ISABEL is that the physician provides accurate source information, which can only be obtained from a careful history and appropriately elicited physical signs. This web-based tool is, therefore, dependent on traditional clinical skills and underlines the importance of the teaching and assessment of effective history taking and examination in medical education. (Greenough 2002, 1259)

As this physician researcher suggests, the role of new tools in the clinic are ones of support. Computer technologies are guided by physicians and their

specialized education and training. The place of AI, thus, should be in service to the clinicians who lead the clinical investigation, take histories, assess the outputs of evidence and deal with the various sources of clinical information. If machines were to have any place in the clinic, then it would be clinicians who determine which technology will be adopted or excluded.

Doctrine

Doctrine, the third element of rarefaction, comprises the modes of reasoning that link individuals within a discourse together (Foucault 1972a, 226). The unifying principles of Western medicine comprise a standardized doctrine that is inculcated and proliferated through medical education. Through training in medical disciplines, doctors develop a form of subjectivity that allows them to interpret evidence, solve problems and enunciate statements in medical terms.

The use of computer technology was increasing across medical education programs throughout the 1980s. Initially, training that relied on information technologies was resisted: ‘new approaches place more emphasis on the use of technology than on the educational mission itself, the transfer and assimilation of information’ (Golden and Friedlander 1987, 851). Resistance was often justified in relation to the cost of equipment, and questions were raised about the effectiveness of machines if students failed to learn to use them correctly. ‘The student should not have the pay more attention to the medium of the message than to the content’ (Golden and Friedlander 1987, 853).

Meanwhile, as ‘black boxes’, AI technologies resist subjection to review; many cannot be understood by analysing their ‘reasoning’ alone. Examining the outputs of neural networks and ML algorithms bears little if any resemblance to evaluating the reasoning of a fellow physician. Thus computers cannot be authoritative enunciators of medical reason, because their decisions cannot be understood: ‘Accuracy and generality aside, *clinicians do expect to understand a decision aid*, wanting insight into the model, its definitions of clinical findings, and its advice. This insight is important if we are to treat such tools as decision aids, not black-box dictators’ (Wyatt 1995, 1176; emphasis added).¹

Social conflict

The fourth and final element of rarefaction concerns the ability to maintain authority over the discourse, those who use it and how it is used. Medical training is central to this endeavour, for ‘every educational system is a political means of maintaining or of modifying the appropriation of discourse, with

¹For an analysis of the difference between interpretation and explanation in relation to the ‘black box’ problem of AI in medicine, see Erasmus, Brunet, and Fischer (forthcoming).

the knowledge and powers it carries with it' (Foucault 1972a, 227). Not only does education establish boundaries, qualifications and access to truth, it also endows the speaker with authority in particular social settings. The words spoken by a trained physician carry a greater force of truth than, for instance, the attending resident or the patient who self-diagnosed on WebMD. In recent years, this struggle for authority over power/knowledge has been especially contentious in AI.

Despite the myth that computers, logic and technical reason are the most direct path to objective knowledge and social progress, recent research has been concerned with implicit bias in the data used to train AI. The struggle for authority in the clinic, first in the 1960s over physicians' subjective judgments (e.g. Sudnow 1967; Hanemaayer 2019b), and now about biases that cannot be checked once programmed into the machine. As one physician writes,

It's well established that clinicians can be influenced by subconscious bias ... AI in healthcare is only as good as the people and data it learns from. This means a lack of diversity in the development of AI models can drastically reduce its effectiveness ... AI trained on biased data will simply amplify that bias. (Novorol 2018)

Justifications prohibiting AI from sharing physicians' clinical authority are extended by the 'black box' problem and the attendant risks of amplified inequalities and marginalization.

The outcomes of AI decision making remain a contentious topic, as evidenced by the promise of IBM's Watson system for oncology and cancer care. Though it was shown to be effective at searching literature about survival rates for treatment options, Watson has never demonstrated effectiveness in treatment recommendations. With Watson's training data now acknowledged to be biased,² it is an open question whether it can operate accurately in diverse clinical settings, such as those in India (Ross and Swetlitz 2017). Citing the negative social outcomes and heavy societal responsibilities that bad decisions carry with them, physicians point to the risk that systems trained on biased data will not, in fact, function well in novel settings or with diverse groups as another reason to resist the encroachment of AI into the clinic.

Discussion: machine learning disrupts the will to truth

This article's discussion of how medical discourse is used in the clinic and by whom provides an empirical picture of the clinic as a site of stratified scientific knowledge. Biotechnology and computer sciences may be actualizing or undermining a shift in the authorship of truth in the clinic, and further research is warranted to explain the antecedent and potential events that may emerge

²Obermeyer et al. (2019) published a scathing study in *Science* about the underlying bias that affects health-related algorithms and AI development. I discuss this later on.

from the intersection of these sciences. To illustrate the stakes, consider Foucault's example of the Penal Code, which first was founded by a theory of Right, but later by the social sciences, such as criminology, psychiatry and sociology (1972a, 219). Similarly, the insertion of biomedical computing and AI into the clinic may be displacing the physicians' responsibility for clinical judgments. In response, the message from medicine appears to be 'don't touch our stuff,' even if only indirectly via justifications specific to their clinical milieu.

This struggle over clinical authority reflects contestation over the will to truth in the clinic, an encounter that takes place through discursive rules that govern 'the principles of classification, ordering and distribution' (Foucault 1972a, 220). The internal rules of medical discourse function by separating true knowledge from other forms of knowledge. For example, sociologists argue that the act of defining 'illness' as 'disease' is premised upon medical classification. For the presenting patient, illness is classified as a disease when it meets certain rules, standards or diagnostic criteria (Jutel 2010, 231). Prior to the application of medical statements, the complaint is merely an illness, unjustified by the functioning of clinical authority.

Machine learning disrupts this function as one that is the physician's alone. Machines are pretty good at classification and have been for quite some time (e.g. Baxt 1995). Their repeated success at diagnosing conditions such as heart attack (e.g. Baxt and Skora 1996) and skin cancer (e.g. Mar, Scolyer, and Long 2017) disrupts clinicians' authority with respect to classification and accuracy.

Who does the classifying raises questions about who is the author of statements. While Foucault (1972a) recognized that the function of authorship may be unifying, the authorship of many significant statements does not matter. When you tell your family you have cancer, they rarely ask the name of the diagnosing physician as evidence of its truth. They will, however, be quite suspicious if it was your aunt who Googled your symptoms or perhaps an AI app to which you uploaded a photo of a skin mole. Likewise, the physicians' statements above all scrutinize *who* is authorized to pronounce truths about the body and its function/dysfunction. While many clinics rely on technologies that have been integrated into the standards of medical care, authority over implementation and test results still remains with the attending physician.

However, by continuously putting pressure on physicians' ability to use sophisticated machines, ML and AI technologies impugn their will to truth and medical authority with it. AI also calls into question physicians' ability to stay abreast of the latest techniques and knowledge, as well as where responsibility for the outcomes of clinical decision making lies. Given the porousness of the clinic to new technologies and novel forms of technical expertise, questions about whose methods and techniques possess medical authority will continue to arise.

Displacing the struggle for truth

In this final section, I raise political questions that may unite multiple actors in a partisanship around new computer technology. I have called Foucauldian analysis ‘political’ insofar as it offers the humanities a counter-human science of the present, one that displaces the human as the object of analysis (Hane-maayer 2018). As I have shown above, the political struggle for authority over truth has been the implicit topic of resistance to and discontent with AI technologies in the clinic. Instead of parodying the clinically discontented as Luddites, would it not be possible to engage in dialogue about that authority and its proper social function?

That conversation might bring together partisan stakeholders who otherwise have very different interests. Physicians and those marginalized by medicine may find themselves in agreement regarding their concerns about the uses of AI for medical classification and diagnosis. Consider the case of racially biased AI systems for patient risk assessment. Obermeyer et al. (2019) found that because a widely used AI algorithm estimated risk using the proxy measure of financial cost, Black patients, who had been allocated fewer economic resources over their medical history despite being biomedically just as sick as their white counterparts, were erroneously assigned a lower risk score, raising the likelihood they would receive sub-optimal treatment (see also Ostherr 2020). Where AI removes the authority of risk assessment from the physician, it risks further marginalizing those who have already been subject to the ills of medical authority. In this context, Ivan Illich (1976) referred to the social effects of medicine as ‘iatrogenic’: the institution of medicine, as he saw it, further entrenched social inequalities. Just as his classic study on the NHS found that populations with the poorest health were not improved by free health care, the use of AI in population health may have similarly unimpressive effects. Against this background, perhaps AI serves as the basis for a new social compact between society and medicine. Could physicians’ struggle against encoded biases unite them with the have-nots and marginalized groups most endangered by those very same biases?

Although this article’s analysis of the history of a forgotten struggle requires further research into the political economy of emerging technologies, their commodification and associated hype to connect them with the myriad institutional pressures from beyond medicine, it is now possible to see what is at stake: who or what society is willing to place trust in when it comes to health and well-being, the body, and its treatment. By elucidating the nature of physicians’ resistance to AI in the clinic, analysing the principles that unify the medical discourse of discontent, and historicizing the struggle over clinical authority in response to computer technologies, this article prompts physicians, patients, biotechnologists and citizens to discuss exactly that: How *should* human and nonhuman activity be organized in the clinic? How should

authority over truth be shared, if at all? What role should non-experts – whether laypeople or abiotic machines – be allowed to play in the clinic?

Is the health of society best served by AI development and computer science? Physicians have suggested over the last decades that it is not. By resisting technical authority and returning to a form of shared decision-making among patients and practitioners, it is possible that the diffusion of discourse in the clinic can be co-authored (c.f. Hanemaayer 2019c).

Acknowledgements

The author would like to thank the reviewers, Shunryu Colin Garvey and Tyler Brunet for comments on earlier drafts of the manuscript. And thanks also due to Tyler Brunet, Adrian Erasmus, Marta Halina and Milena Ivanova for giving me the idea to write this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Ariane Hanemaayer is an associate professor at Brandon University and a visiting fellow at the Centre for Research in Arts, Social Sciences and Humanities at the University of Cambridge.

ORCID

Ariane Hanemaayer  <http://orcid.org/0000-0001-6921-6887>

References

- Baxt, William G. 1995. "Application of Artificial Neural Network to Clinical Medicine." *The Lancet* 346: 1135–1138.
- Baxt, William G., and Joyce Skora. 1996. "Prospective Validation of Artificial Neural Network Trained to Identify Acute Myocardial Infarction." *The Lancet* 347: 12–15.
- Bayram, Mustafa, Simon Springer, Colin K. Garvey, and Vural Özdemir. 2020. "COVID-19 Digital Health Innovation Policy: A Portal to Alternative Futures in the Making." *OMICS: A Journal of Integrative Biology* 24. doi:10.1089/omi.2020.0089.
- Beautyman, William. 1982. "Natural Calculators, Polydactyly, and the Diagnostic Process." *The Lancet* October 23: 932.
- Benítez-Bribiesca, Luis. 2000. "In Defense of the Clinical Art." *Lancet* 355: 760–761.
- Campbell, E. J. M. 1987. "The Diagnosing Mind." *The Lancet* April 11: 849–851.
- Dodds, R. S. 1995. "Neural Networks." *The Lancet* 346: 1500–1501.
- Erasmus, Adrian, Tyler Brunet, and Eyal Fischer. forthcoming. "What Is Interpretability?" *Philosophy & Technology*. doi:10.1007/s13347-020-00435-2.
- Foucault, Michel. 1972a. "Appendix: Discourse on Language." In *The Archaeology of Knowledge*, 215–237. New York: Harper & Row, Publishers.
- Foucault, Michel. 1972b. *The Archaeology of Knowledge*. New York: Routledge.

- Foucault, Michel. 1998. "Nietzsche, Genealogy, History." In *Michel Foucault: Aesthetics, Method, and Epistemology*, vol. 2 of *Essential Works of Foucault 1954–1984*, trans. Robert Hurley et al., edited by James D. Faubion, 370–381. New York: New Press.
- Golden, William E., and Ira R. Friedlander. 1987. "Inverse Technology and Medical Education." *Lancet* April 11: 851–853.
- Goldhahn, Jorg, Vanessa Rampton, and Giatgen A. Spinus. 2018. "Could Artificial Intelligence Make Doctors Obsolete?" *BMJ* 363: k4563–k4565. doi:10.1136/bmj.k4563.
- Greenough, Anne. 2002. "Help from ISABEL for Paediatric Diagnoses." *The Lancet* 360: 1259.
- Hanemaayer, Ariane. 2018. "Genealogy as a Counter-Human Science." *Canadian Review of Sociology*, doi:10.1111/cars.12197.
- Hanemaayer, Ariane. 2019a. "Doing Archival Research." *Craft of Qualitative Inquiry*. Toronto, ON: Canadian Scholars Press.
- Hanemaayer, Ariane. 2019b. *The Impossible Clinic: A Critical Sociology of Evidence-Based Medicine*. Vancouver: University of British Columbia Press.
- Hanemaayer, Ariane. 2019c. "The Ethic of Responsibility: Max Weber's *Verstehen* and Shared Decision-Making in Patient-Centred Care." *Journal of Medical Humanities*, DOI: 10.1007/s10912-019-09577-7.
- Horton, Richard. 2019. "Offline: Touch – the First Language." *The Lancet* 394: 1310.
- Illich, Ivan. 1976. *Medical Nemesis*. New York: Pantheon Books.
- Ioannidis, John P.A., and Joseph Lau. 1996. "On Meta-Analyses of Meta-Analyses." *The Lancet* 348: 756.
- Jutel, Annemarie. 2010. "Medically Unexplained Symptoms and the Disease Label." *Social Theory & Health* 8 (3): 229–245.
- Kickbusch, Ilona, Anurag Agrawal, Andrew Jack, Naomi Lee, and Richard Horton. 2019. "Governing Health Futures 2030: Growing up in a Digital World – a Joint *The Lancet* and *Financial Times* Commission." *The Lancet*, doi:10.1016/S0140-6736(19)32181-6.
- Lindberg, Donald A. B. 1990. "In Praise of Computing." In *A History of Medical Informatics*, edited by Bruce Blum, and Karen A. Duncan, 4–11. New York: ACM Press.
- Mar, Victoria J., Richard A. Scolyer, and Georgina V. Long. 2017. "Computer-Assisted Diagnosis for Skin Cancer: Have We Been Outsmarted?" *The Lancet* 389: 1962–1964.
- Novorol, Claire. 2018. "How AI Can Develop Bias and Discriminate Against Patients." *AI in Medicine*. Accessed August 12, 2020. <https://ai-med.io/ai-med-news/ai-biases-ada-health-diversity-women/>.
- Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–453.
- Osborne, Thomas. 1999. "The Ordinarity of the Archive." *History of the Human Sciences* 12 (2): 51–64.
- Osther, Kirsten. 2020. "Artificial Intelligence and Medical Humanities." *Journal of Medical Humanities*, doi:10.1007/s10912-020-09636-4.
- Ross, Casey, and Ike Swetlitz. 2017. "IBM itched its Watson supercomputer as a revolution in cancer care. It's nowhere close." *Stat*. Accessed August 14, 2020 <https://www.statnews.com/2017/09/05/watson-ibm-cancer/>.
- Signorini, David F., and Jim M. Slattery. 1995. "Neural Networks." *The Lancet* 346: 1500.
- Sudnow, David. 1967. "Dead on Arrival." *Trans-action* 5: 36–43.
- Taylor, Thomas R. 1967. *The Principles of Medical Computing*. Oxford: Blackwell Publishing Ltd.
- Walters, William. 2012. *Governmentality: Critical Encounters*. London: Routledge.
- Wyatt, Jeremy. 1995. "Nervous About Artificial Neural Networks?" *The Lancet* 346: 1175–1176.



Clinical translation of computational brain models: understanding the salience of trust in clinician–researcher relationships

S. Datta Burton ^a, T. Mahfoud^b, C. Aicardi ^c and N. Rose^c

^aDepartment of Science, Technology, Engineering and Public Policy (Steapp), University College London, London, UK; ^bDepartment of Sociology, University of Essex, Colchester, UK; ^cDepartment of Global Health and Social Medicine, King's College London, London, UK

ABSTRACT

Computational brain models use machine learning, algorithms and statistical models to harness big data for delivering disease-specific diagnosis or prognosis for individuals. While intended to support clinical decision-making, their translation into clinical practice remains challenging despite efforts to improve implementation through training clinicians and clinical staff in their use and benefits. Drawing on the specific case of neurology, we argue that existing implementation efforts are insufficient for the responsible translation of computational models. Our research based on a collective seven-year engagement with the Human Brain Project, participant observation at workshops and conferences, and expert interviews, suggests that relationships of trust between clinicians and researchers (modellers, data scientists) are essential to the meaningful translation of computational models. In particular, efforts to increase model transparency, strengthen upstream collaboration, and integrate clinicians' perspectives and tacit knowledge have the potential to reinforce trust building and increase translation of technologies that are beneficial to patients.

KEYWORDS

Human Brain Project; Artificial Intelligence; machine learning; neurology; neuroscience; clinical prediction models; big data; data driven health

1. Introduction

At the 2018 Annual Summit meeting of the Human Brain Project (HBP), the chair of the short-lived Clinical Advisory Board pointed out that,

when you go see your doctor, you must feel confident in their diagnostics, be reassured, trust them. The paradox is, currently doctors don't or can't, capitalise on the previous hundreds of patients them and their colleagues may have seen, who may have displayed similar symptoms.

At the same time, the chair urged researchers to remember that when developing even the most wonderful of medical tools to fix this problem with Big Data

CONTACT S. Datta Burton  saheli.burton@ucl.ac.uk  Department of Science, Technology, Engineering and Public Policy (UCL STEaPP), University College London, Shropshire House (4th Floor), 11–20 Capper Street, WC1E 6JA, London, UK

© 2020 Institute of Materials, Minerals and Mining Published by Taylor & Francis on behalf of the Institute

analytics, it is always mandatory to develop ‘the human side of these tools’. Where data-intensive algorithms for brain disorders are concerned, this ‘human side’ represents a triangle of trust relationships: between patients and researchers,¹ patients and clinicians, as well as clinicians and researchers. In this article, we explore questions of trust between patients and clinicians and between clinicians and researchers. As we shall see, these questions have implications for the use of diagnostic technologies based on computational modelling and machine learning techniques for the analysis of the vast quantities of patient data gathered by hospitals from clinical research and elsewhere.

In the HBP, work is underway to analyse large clinical and research datasets, using algorithms and machine learning. The aim is to search for patterns in data that can individuate the neurobiological correlates of a disorder in ways that could be used to aid diagnosis, to target treatments, and hence to improve prognosis. Computational models are one among many types of approaches currently under development to integrate Artificial Intelligence (AI), broadly construed, in healthcare. In turn, these computational models draw on the vast amount of personal health data now available to develop disease-specific machine learning algorithms and statistical models that are expected to assist clinical decision-making when offering a ‘personalized’ diagnosis or prognosis for individuals.

Recent surveys and mappings of the ethical and social questions raised by the use of AI in healthcare² place overwhelming emphasis on issues of *public trust*, hinged on questions of bias and explainability of algorithms (Watson et al. 2019), as well as transparent, fair, secure and equitable use of health data (Future Advocacy 2018; Joshi and Morley 2019). Epistemic concerns with AI applications for healthcare remain mostly related to evidence as inconclusive, inscrutable, or misguided (Morley et al. 2020; Wessler et al. 2017). Stakeholders engaged with improving implementation (and thus uptake) of AI applications remain focussed on training as a key tool for overcoming clinician’s assumed distrust (Liberati et al. 2017) of and reluctance to learn novel techniques (Future Advocacy 2018; Joshi and Morley 2019; Nuffield Council on Bioethics 2018).

Yet this focus on issues of public trust, evidence, and training misses and obscures the salient role of tacit knowledge in clinical diagnosis (Allegaert, Smits, and Johannes 2012) emblematic of the long, and ongoing, struggle of clinical practice to establish its epistemic value on firm and widely accepted grounds (Khushf 2013; Malterud 1995; Leblond 2013; Engel 2008). This struggle is inseparable from clinician and clinical staff’s fears of job loss

¹In this paper, ‘researchers’ refer to modellers neuroscientists and data scientists involved with the development of computational models and data analysis tools.

²Covering all areas of use from process optimisation to patient-facing applications and notably applications integrated in clinical pathways, which encompass among other techniques machine learning and data-driven computational models.

contingent on perceived threats to autonomy, authority and expertise, loosened relations with patients, deskilling, jobs displacement (Greenhalgh et al. 2017; Simonite 2016; Rockoff 2016). We suggest understanding and integrating these entangled and contingent complexities of the clinician's perspective and the contribution of tacit knowledge towards building meaningful relationships of trust between clinicians and researchers as a first step in efforts to routinize computational brain models in neurological practice (see Figure 1). For patients (and thereby publics) need to trust clinicians as the custodians of their welfare, and by and large do so. Thus we argue that inadequate attention to the salience of clinician's trust in computational models risks weakening efforts to (re)gain public trust in AI-driven healthcare.

We focus here not on patients' or publics' trust, but on a different and equally important dimension: *clinicians' trust*; a key factor in research–practice partnerships vital to the translation of machine learning and algorithm-driven technologies. We reflect on the role of clinicians both as custodians of patient trust and welfare, and as end-users of computational models. In turn, this highlights the challenges and barriers to the responsible and successful adoption of such technologies; the clash between epistemic cultures and professional practices of data science and medicine; and the implications these have for how data are gathered and interpreted. Overcoming such challenges, we suggest, will depend on the trust of clinicians that their tacit, experiential, clinical knowledge is respected and integrated into data-driven technologies and that these technologies will meaningfully benefit patients.

We begin by briefly discussing some definitional aspects of trust in the context of the 'patient–clinician' relationship before considering the ways that clinician's trust in computational models in neurology shapes, and is shaped by, their traditional interpersonal trust relationship with patients. This is followed by an analysis of the implications of clinician's relationship with researchers. This line of inquiry extends scholarship in healthcare and biomedicine around methodological concerns such as data anonymisation (Watson et al. 2019), consent (Larson 2013), platform standardization (Shah, Steyerberg, and Kent 2018) etc. At the same time, our inquiry provides a unique qualitative understanding of the ways in which trust relationships shape (or weaken) patient trust (Hall et al. 2002;

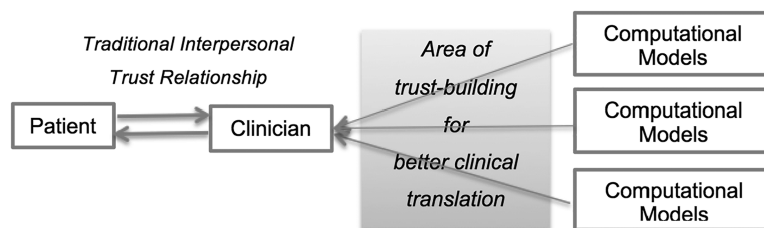


Figure 1. Trust relationships in clinical translation of CPMs.

Fiscella et al. 2004; Thom 2001; Thorpe et al. 2020; Klein et al. 2016). The empirical research for this article is grounded in our collective seven-year engagement with HBP as researchers of its ‘Ethics and Society’ subproject and includes analysis of published and grey literature, participant observation at workshops and conferences, and interviews with data scientists, neuroscientists, and neurologists in the UK and Europe developing computational tools for neurology.

2. Background

2.1. Traditional clinician–patient trust relationship

Trust in science, medicine, and experts in general has been the subject of much academic and popular debate since at least the 1960s, although a general definition of trust remains elusive and contested (see exhaustive discussion in Mcknight and Chervany 1996). In contrast, specific ‘trust’ relationships like the patient-clinician trust relationship have been widely studied (Hall et al. 2002; Fiscella et al. 2004; Thom 2001; Thorpe et al. 2020; Klein et al. 2016). However, most studies have used large-scale survey data to identify objective measurable conditions that erode or enhance patient trust such as ‘perceived clinician financial conflicts of interest’ has been shown to erode trust (Klein et al. 2016). While patient beliefs of clinician’s honesty and competence to ‘act in their [the patient’s] best interest, and preserve their confidentiality’ (Fiscella et al. 2004), clinician’s efforts to understand patient experiences and share power (Thom 2001) etc. have been found to enhance patient trust in clinicians.

Yet, whether measuring the erosion or enhancement of trust, these studies – perhaps more than anything else – implicitly emphasize the enduring nature of the interpersonal relationship of trust that patients have with clinicians as custodians of their welfare. A key reason for this, according to 70% respondents of a nationwide survey of 3014 US adults by Pew Research, was because clinicians were ‘the central resource for information or support [for patients, carers, and family] during serious episodes’ and at other times (Fox and Duggan 2013). As to ‘why do individuals trust their doctors the most?’, the answer (according to another large survey conducted by PricewaterhouseCoopers 2012) was ‘human relationships’. Hence, when novel clinical technologies are introduced, patient’s (and thereby publics) decisions to adopt or reject them are overwhelmingly informed and guided by advice from clinicians. This is because ‘[patients (and publics)] want to trust and connect with the people providing [them] the care. ... it’s easier to trust a person than an organization ... [and clinicians] have the ability to form human relationships and connections with their patients, which ultimately leads to increased trust’ (Kathryn Armstrong, senior producer of web communications at Lehigh Valley Health Network, USA in PricewaterhouseCoopers 2012, 17). Indeed, the salience of this trust relationship is highlighted by PricewaterhouseCoopers’ (2012, 17) survey of

patient behaviours (aimed at understanding the nuances of health technology adoption) which concluded on a cautionary note that ‘to establish trust and credibility with consumers ... healthcare companies need[ed] to reconsider their approach to these [clinician–patient] relationships’.

Within this complex interpersonal patient–clinician trust relationship, influenced to some extent by patient’s institutional trust of the hospital or clinic where the clinician is embedded and from where they receive health (care) services (Gray 1997), computational models hope to gain a foothold as a trusted member of the clinician’s diagnostic toolkit alongside stethoscopes and blood pressure monitors. Next we turn to the case of neurology where these models are being developed to aid neurologists in order to to understand the challenges of how computational tools can gain a foothold - in other words, gain the clinician’s trust.

2.2. Neuro-diagnosis

Neurology has a long history of attempts to codify the diagnostic process but with limited success, especially around the interpretation of medical imaging, and formalizing it in standard and computerized programmes (Doi 2007). Enormous research efforts over several decades, involving genetic, scanning and other advanced neurotechnologies attempted to identify neurological ‘bio-markers’ for any of the current diagnostic categories used in clinical practice, for example, those embodied in the successive editions of the American Psychiatric Association’s *Diagnostic and Statistical Manual of Mental Disorders* (Rose 2013). However, with the exception of some forms of dementia, these attempts failed to identify clinically useful neurobiological or genetic bio-markers for diagnostic precision or treatment choice in the area of mental health. This led many psychiatric researchers, notably at the US National Institutes of Mental Health, to argue for a shift away from research based on diagnostic categories towards developing new approaches that would diagnose disorders on the basis of their biology.

The hope was that novel machine learning techniques when used to analyse large data sets (containing information from genetic tests, brain scans, and other physiological markers, together with data on clinical presentation, symptomatology, treatment, and prognosis) might reveal previously hidden relations between neurobiology, symptomatology, and treatment success. Computational models are emblematic of such attempts to formalize and standardize the diagnostic process. By codifying the range of data now available for neuro-diagnosis from various sources (fMRI, MRI, PET, CT, EEG, MEG), models aim to develop ‘objective’ diagnoses, rather than person-driven analyses dependent on the interpretive skills of different clinicians. Currently, such approaches are being tried across a number of areas, to analyse case records, clinical and physiological data, test results and images from scans and link these to diagnosis and

prognosis – in some cases producing results that are more accurate and reliable than those of even the most skilled diagnostician (Nuffield Council on Bioethics 2018). Not surprisingly, computational brain models have become an area of considerable commercial investment, bolstered to some extent – as in the United States – by the hope of automating clinical decision-making to reduce diagnostic errors held responsible for the rising cost of medical malpractice settlements.³

2.3. Tacit knowledge

For many scholars, the challenge lies in trusting the ‘reading’ of data produced by these computational models. For instance, many have recognized that images (medical imaging data) do not speak for themselves, are not mere extensions of the naked eye and can only assist, not replace, expert opinion (Stone et al. 2016). As clinicians such as Barry F Saunders (2008) demonstrated, the ‘craft’ practices involved in ‘learning to read’ radiological images such as CT scans encapsulate complex institutional and hierarchical context within which doctors are trained to develop ‘tacit knowledge’ (see detailed discussion in Mahfoud 2014). Anthropologist Andreas Roepstorff (2007) called this ‘skilled vision’ – or ways of knowing that cannot easily be described, explained or put into words, but which enable us to do what we do (Polanyi 2009) – in this case to draw conclusions from visual and other evidence to make a diagnosis that will lead to a decision about action.

Unlike, earlier neurologists such as Kurt Goldstein (1878–1965) who had access only to the brains of deceased patients, and thus diagnosed living patients using case records and visual technologies such as films (Goldstein 1995[1939]) neurologists today have access to the living brain via a range of imaging technologies for diagnostic purposes – CT, PET, MRI, fMRI, EEG, etc. (Rose and Gainty 2019). These images are generated by sophisticated technology and statistical procedures whose details are often unknown to those who make use of them. For instance, magnetic resonance imaging (MRI) uses magnetic field gradients that act upon certain atoms to generate detectable radio waves that are then processed using sophisticated algorithms to generate data on the distribution of water and fat in the body and then further processed to generate images of organs. However, these images embody many assumptions. For example, the hypothesis built into fMRI is that an increase in blood flow is a marker of an increase in brain activity. This assumption is embedded in the software embedded in the fMRI scanner that produces the images and despite longstanding critical evaluation, it is seldom questioned in the practices that utilize fMRI for research or diagnosis (Logothetis 2008). Indeed the

³see analysis by researchers at Johns Hopkins https://www.hopkinsmedicine.org/news/media/releases/diagnostic_errors_more_common_costly_and_harmful_than_treatment_mistakes

standard fMRI paradigms have become controversial with the increasing recognition that measures of changes in blood oxygenation levels neglect the key role of highly distributed neural activity – known as the ‘resting state’ – that is necessary for task performance, but is usually ‘subtracted’ from fMRI outputs as a result of the algorithms that are used to create the images (Gusnard, Marcus, and Raichle 2001).

These emerging data-driven technologies are thus far more than an aid to vision; they render some things visible at the expense of others, and frequently do so in ways that are ‘black boxed’ and not known or fully understood by those who use the results (Caruana et al. 2020; Holzinger et al. 2017). These issues are further complicated when large volumes of data produced by these technologies in many different clinics and research projects, using different research protocols, are linked using statistical devices to make them commensurable, and then analysed using machine learning to generate algorithms that are not known, let alone fully understood by potential users (the clinicians) who would have to decide whether and when to make use of the information provided in order to make a medical diagnosis.

Thus, many argue that as in previous diagnostic practices, ‘reading’ these images require the intervention of the trained eye of the expert (Mahfoud 2014), specifically trained in tacit practices typically taught outside of formalized education via apprenticeships (revealed in interviews with neurologists; see also Shah, Steyerberg, and Kent 2018). As a result, the interpretive skills developed by neuroscientists (e.g. to interpret fMRI brain scans) involve both formal and informal education – one is required to know in order to see (Roepstorff 2007). In the clinical practice of neurology, ‘knowing’ (in order to read images) crucially draws on clinician’s tacit knowledge of individual patient’s physiologies and pathologies gained through interpersonal clinician–patient trust relationships. Indeed, the foundational salience of tacit knowledge in psychiatry is revealed by a recent survey of 791 psychiatrists across 22 countries (representing North and South America, Europe and Asia-Pacific) where a mere 3.8% ‘felt it was likely that future technology would make their jobs obsolete and only 17% felt that future AI/ML was likely to replace a human clinician for providing empathetic care’ (Doraiswamy, Blease, and Bodner 2020) (see also Miner et al. 2019). As Ferdinand Velasco, Texas Health’s chief medical information officer (op. cit 27) emphasizes,

There is a lot of patient data – clinical and soon genomics as well. But what is really happening in our patients’ lives is missing to us and their record– what’s happening in their lives is happening in the social space. ... If we understand the life factors that impact when and who they select for care and what challenges they face after receiving care, there is a lot of potential for merging analytics with the clinical side and improving care. The lack of access to various facets of human suffering that lies underneath the bare data is particularly acute for organisations and institutions engaged in

translational research where researchers traditionally have no direct interaction with patients except during clinical trials. (PricewaterhouseCoopers 2012, 17)

Thus, many argue that engagement is crucial between researchers and clinicians early on in technology development to bridge the widening ‘chasm in the understanding of end-users [clinicians] between [researcher’s] imagination of a future user and users’ lived experiences’ (Datta 2018, 354; see also Epstein 1996; Smith, Bossen, and Kanstrup 2017). Without this early engagement, the expectation that once a technology is developed it will be readily adopted by clinicians (and patients) – at the researchers’ word of its power to improve their lives – is challenging. While engagement efforts after a technology are developed is widely considered a meaningless tick-box approach that is ‘largely ineffective in rebuilding public trust’ (Wynne 2006, 217). Thus, from the clinician’s perspective, the computations involved in producing these images in order to ‘read’ and analyse brain scans hold little practical interest. However, as we shall see, clinicians do need to have confidence and trust in the process that has led from the initial data to the images that they have to interpret and utilize in clinical practice.

2.4. Computational brain models

In contrast, from the perspective of the researchers seeking to interpret the range of information potentially available for diagnosis, machine learning promises to solve the problem of the masses of data from different sources now available – fMRI, MRI, PET, CT, EEG, MEG ... – that are extremely difficult for human beings to integrate and analyse together. The hope is that machine learning tools can be used to analyse these large amounts of data from different sources, and further, that these computer-driven methods will be more objective than person-driven analyses, because they do not depend on the interpretive skills of different clinicians. Thus many argue that the digitization of data from patients’ clinical records and medical images combined with advanced data analytics can enable AI technologies such as machine learning and machine vision to distinguish between different potential diagnoses in a clinically meaningful way that can enable clinicians to target specific treatments (Luo et al. 2016). Data-driven technologies aimed at assisting healthcare practitioners in diagnostic processes have been approved by the US Food and Drug Administration (FDA) in recent years (Future Advocacy 2018). For example, the smartphone application ‘Viz.AI’ analyses CT images of the brains of patients admitted to hospitals with symptoms of stroke, identifies vessel blockages through these images, and sends this analysis via text to neurovascular specialists. This software was approved by the FDA as a ‘clinical decision support software’ based on evidence submitted by the developers which demonstrated through a clinical trial that the software application more

quickly identified the vessel blockages.⁴ Development of several other software applications is underway (although not yet approved by the FDA), such as the collaboration between DeepMind Health and Moorfields Eye Hospital, London which uses neural networks to diagnose Age-related Macular Degeneration (AMD) through the analysis of Optical Coherence Tomography (OCT) (De Fauw et al. 2018).

In psychiatry, much of the research still begins from or utilizes diagnoses according to current diagnostic categories despite the problems that many have identified with such classification (discussed earlier). Thus a collaboration between IBM and the University of Alberta uses neural networks to diagnose schizophrenia through an analysis of fMRI scans while patients undertake an audio-based exercise. The researchers claim to have identified ‘combinations of statistical features extracted from the data that can serve as reliable statistical (bio)markers of the disease, capable of accurately discriminating between schizophrenic patients and controls’ (Gheiratmand et al. 2017). These ‘bio-markers’ included an ‘abnormal’ increase of connectivity between the thalamus and the primary motor/primary sensor cortex as well as ‘hyperconnectivity’ in the fronto-parietal network. While the model was relatively successful at predicting a clinical diagnosis of schizophrenia (at above 70%), this research has not yet undergone the clinical trials needed for software regulation and approval. Further, it is being undertaken at a time when the very categories such as schizophrenia are contested by many experts in the field (Murray 2017). There is a widespread recognition that, across the whole spectrum of mental disorders, and particularly in relation to psychoses, similar symptomatology may result from very different neurobiological pathways. There is thus a certain unhelpful circularity in seeking brain based biomarkers that correlate with symptom based diagnoses that are themselves contested and considered to lump together a variety of conditions that are developmentally, neurobiologically, and prognostically distinct (Nature Biotechnology 2012).

Some six years ago, an Editorial in *Nature Biotechnology* (2012) entitled ‘What happened to personalized medicine?’ reflected on the slow progress and unrealized hopes of those who predicted a revolution in medical diagnosis and treatment targeting based on biomarkers. The barriers they identified were less biological than social: a need ‘to broaden the concept of personalized medicine from the genetically reductionist version to one that includes other types of markers’; a need for more long-term studies ‘linking specimens, sequence and other biomarker information to clinical outcomes’, a need for patients to be encouraged to share their data for research purposes, and a need to educate physicians ‘about the new diagnostics and how to integrate them with existing clinical information’ which will require not only better education but also ‘the development of robust point-of-care devices and data-sharing technology and

⁴<https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm596575.htm>

the establishment of trusted sources' (e.g. medical association position statements on tests or the National Institutes of Health's genetic testing registry). What perhaps stands out most in the editorial, is its emphasis on the salience of trust. For if clinicians and patients do not have legitimate trust in the accuracy, validity and utility of biomarkers, and that certainly includes brain based biomarkers, whatever the hopes of those who develop them, they will not 'translate' into clinical practice.

3. Gaining the clinician's trust

So far, the literature on the dynamics of gaining the *clinician's trust* in clinical translation mostly derive from research focussed on other issues such as studies of computational technology adoption rates among clinical staff (Garland, Plemmons, and Koontz 2006), or critical scholarship identifying barriers to effective research–practice partnerships for better clinical translation or data collection (Mittelstadt and Floridi 2016). In turn, scholarship on barriers to the adoption of computational models draws on these views of resistance to change in research–practice relationships to implicate this lack of trust as one among several 'technical and methodological' issues such as calibration, risk-sensitivity, data quality (Shah, Steyerberg, and Kent 2018), and so forth. However, some research does acknowledge the need to contextualize these understandings within the imperatives of big-data processing (Mittelstadt and Floridi 2016, 2–3).

3.1. Upstream clinician collaboration

Our research shows that collaborations with clinicians in the upstream research conceptualization phase of technology development were crucial for gaining the clinicians' trust, as it provided the time and space to forge an interpersonal researcher-clinician trust relationship necessary to create bridges across the researcher-clinician divide. Typically, the relationship between clinicians and computational neuroscientists is based on give and take – clinic(ians) send researchers anonymized patient data, researchers perform analyses using this data in computational models under development and send back the results of the analyses to clinic(ians). However, our interviews revealed that trust was a defining element in this relationship between clinicians and researchers, particularly around data bias, model transparency, and different epistemological traditions in neurology, the neurosciences, and computer sciences.

Importantly, researcher-developers of neuro-diagnostic tools highlighted the significance of close collaborations between clinicians, or 'domain experts', and data analysts, not just for the sharing of medical data but more importantly for defining the research questions pursued. Computer scientists stated that research questions should be defined by clinicians, and only then can machine learning tools or methods be selected and decisions made about

which types of data are needed. Similarly, there was acknowledgement among modellers that ‘domain-specific’ training would be desirable to collaborate with clinicians, beyond the generalist traditions of computer science that require computational tools be developed for general-purpose and then adapted to specific use-cases. In reality, while some modellers do develop expertise in specific fields (such as neuroscience, or oncology), it is more common for computer scientists and engineers to move between biological domains of expertise.

Clinicians and biologists, on the other hand, found it concerning that modellers did not understand the biology and physiology of the conditions being explored, which they deemed necessary to develop clinically useful models. At the same time, researchers found it concerning that clinicians did not understand the modelling frameworks used – specifically the assumptions of the statistical, machine learning and other data analysis methods. Computational modellers suggested that future educational programmes for clinicians need to include training in computational methods in order to adapt to changing clinical contexts where machine learning and other computational tools are likely to become more commonplace.

3.2. Greater model transparency

For clinicians, model transparency played a considerable role in deciding the extent of their trust in using a machine-learning-based diagnostic tool. Some researchers argue for shorter-term technical solutions such as making machine learning tools more interpretable for clinicians by building in the ability for clinicians to trace back the way an algorithm has come to a certain conclusion, thus rendering the decision-making ‘transparent’. Indeed, transparency is one of the key ethical principles in discussions around accountability in artificial intelligence (Mittelstadt and Floridi 2016). An association of researchers in Microsoft, Google, and others have, for example, proposed the principles of ‘Fairness, Accountability, and Transparency in Machine Learning’ to address the ‘potentially discriminatory impact of machine learning’ as well as the ‘dangers of inadvertently encoding bias into automated decisions’.⁵ Interviews with computational modellers suggested that to adhere to the principle of transparency, the use of supervised and semi-supervised learning algorithms was preferable to relying on unsupervised learning algorithms. This is because supervised and semi-supervised classification algorithms can be represented as decision-trees, which are more interpretable to collaborating clinicians than the ‘black box’ through which unsupervised learning algorithms produce their results.

Transparency and open science are core values of responsibility in scientific research and innovation (Von Schomberg 2013). These issues have become

⁵<http://www.fatml.org/>

prominent in debates over algorithmics, with the demand for explainability and accountability of inscrutable systems getting stronger. This demand is gaining traction as it is increasingly believed to be a crucial, perhaps inescapable, step towards safety and trustworthiness of AI and machine learning systems, and it is seen as integral to ethically aligned design principles. Unsupervised learning algorithms, which aim is to discover inherent structures in data without using pre-existing categories, are under special examination for being inscrutable even to their designers. In this context, any proposals to use such unsupervised machine learning to discover ‘brain signatures’ that could bring about a complete revision of the classification of mental disorders appears highly problematic.

3.3. Integrating tacit knowledge

Even if algorithms can become less opaque, there remain other epistemological obstacles – that is to say, clashes of epistemologies – to collaborations between clinicians and researchers. Clinicians we interviewed talked about the importance of clinician-patient interaction for diagnosis. For example, in the diagnosis and treatment of epilepsy, clinicians carry the patients through from diagnosis to pre-surgical screening, surgery, and post-surgical rehabilitation. This is seen by some as already a highly personalized treatment since each patient is unique in terms of the symptoms exhibited and surgical treatment needed. Such ‘holistic’ treatment of the individual – common in neurology – is seen by computational neuroscientists as dependent on ‘subjective’ and ‘biased’ elements that need to be removed or reduced from clinical settings described as ‘low validity environments’ with some citing psychologist Daniel Kahneman’s (2013) work on decision-making: ‘to maximize predictive accuracy ... decisions should be left to algorithms in low validity environments’ for support.

Despite such criticisms, ‘tacit knowledge’ of clinicians remains crucial in diagnostic processes. For example, it is clinicians who must use their formal and informal craft skills to interpret the various data sets from the patients to decide which part of the brain to remove during surgery or where to place SEEG (stereo electro encephalography) electrodes for an epileptic patient. In response to a question about the receptivity among clinicians of computational brain modelling approaches for the diagnosis and treatment of neurological conditions, an engineer at a neuroscience laboratory in France said:

Clinicians do not explicitly state the idea behind why brain areas generate seizures, for example – it is an implicit model. I call it a model. It is what you learn when you study epileptology. Clinicians reason by saying that if this area affects this one, then if I remove this it will stop seizures. It is a model because it is a set of rules, but it is not entirely quantitative. There is a lot of experience needed to know, to have intuition. The goal of our quantitative model is to take into account the clinician’s opinion but also other data to come up with results that bring those things together, hopefully

in a robust fashion. The clinicians we work with are interested in having tools that will help them resolve the pathology for their patients in cases where they have no idea ... They will try [the data analysis] out because they are curious, but will ignore the result. But as they get experience with the tool, assuming it is good enough, it will become part of the workflow and part of their analysis.

For this researcher as well as others who we interviewed, such tacit or experiential knowledge is an implicit model as it involves a series of steps that result in a decision, or solution; they thus expressed concern that these steps were not made explicit. On the other hand, clinicians defended this ‘intuition’ as the result of years of diagnostic experience and surgical interventions that become internalized in their judgement processes. Clinicians insist that just because these were not fully articulated, or perhaps even capable of full articulation, does not mean they were not valid. While they did not use the term, it was clear here that clinicians considered that a ‘skilled clinician’ was one who had fully embodied the results of years of experience and immersion in a community of experts into the tacit knowledge that guided their decisions.

Not surprisingly, this epistemological conflict between the ‘subjective’ knowledge of the clinicians and the so-called ‘objective’ knowledge of the data analysis finds its way into discussions around data bias. Data scientists need what they call ‘good data distribution’, standardized data, and comprehensive meta-data. As one computer scientist noted:

There are different cultures of how clinicians diagnose and treat – different clinicians with different expertise and knowledge can label patients differently. And many scores – like the Montreal Cognitive Assessment (MoCA) – are arbitrary, on a scale of one to five or something. This kind of noise can be compensated for if you have large amounts of physicians involved who can make sure the data is distributed well and can check the quality of measurements, and missing data.

On the one hand, this suggests that including clinicians’ tacit knowledge meaningfully in upstream technology development processes – rather than as part of a box-ticking exercise in post-development processes – is important for meaningful technology adaptation and its eventual adoption (Wong, Turner, and Yee 2008). This is consistent with Sullivan et al.’s (2005) conclusion in the domain of psychiatry research and practice, that a bottom-up ‘approach in which services researchers assist frontline clinicians in testing interventions that clinicians themselves have devised ... [will] result in interventions that are more likely to be sustained over time’. For instance, recent work on ‘automatic speech recognition for psychotherapy’ – whereby nuances of patient’s speech in datasets used to develop a model vary significantly from patient populations it hopes to serve – highlight the acute need for clinician engagement in designing and guiding model development (Miner et al. 2020) to reduce clinical bias in existing data and better reflect variability’s in race, ethnicity, sex, gender, age, etc. (Schwartz and Blankenship 2014).

On the other hand, modellers and clinicians alike acknowledge that clinicians do not have the required training to critically analyse the results of these new data analytics tools, and their distrust of these results may be caused by a lack of understanding of the models and the ways algorithms reach decisions. While unsupervised learning has been successful in diagnosis, especially in relation to the analysis of medical images (e.g. by Google DeepMind in De Fauw et al. 2018), these are far less interpretable than other machine learning methods. Yet, whether interpretable or not, clinicians we interviewed for this study suggested that if the computational models recommended a different diagnosis or different conclusions than their medical opinion, they were almost always likely to trust and pursue their own opinion.

For computational diagnostics tools aiming to become a part of the neurologist's toolkit, there is a need to move away from the negative 'quantitative' view that 'subjectivity' of clinicians' assessments are inherently 'biased' and acknowledge the value of experiential and tacit knowledge in clinical reasoning that remains foundational to the relationships of trust between patients and clinicians. Incorporating 'skilled vision' within algorithms (Roepstorff 2007) is potentially the first step towards overcoming the difficulties of AI routinization into practice for those technologies that are perceived to add value to clinical reasoning rather than competing with it (like the new generation of AI-integrated or 'smart' computation diagnostics tools) are more likely to win clinician's trust.

Following this, the next step, as recent work in psychiatry suggests, is the need for careful research into the qualitative nuances (human impacts) of deploying 'AI delivered, human [clinician] supervised psychotherapy' (Bhugra et al. 2017; Miner et al. 2019; Patel et al. 2018). A priority for such deployments of AI-enabled care must be to avoid situations where patient loss of trust in AI adversely reflects on patient trust in clinicians and in care provisioning (Miner et al. 2019; see also Bhugra et al. 2017; Patel et al. 2018). This is especially likely if ever higher 'expectations of benefit [from ever more sophisticated AI]' cause patients to 'transition from feeling let down to feeling betrayed [by clinicians and by the systems of "care"]' (Miner et al. 2019).

3.4. From clinician's trust to investor trust

For private investment to support the long process of clinical translation to the market, the technology end-user's (here the clinician) trust in and willingness to adopt the new technology is crucial (Greenhalgh et al. 2017). For instance, research shows that having clinicians in management positions with computational skills is positively correlated with 'long-term commitment to the use of IT [in health care]' (Ingebrigtsen et al. 2014). The reverse – clinical leadership lacking a commitment to IT-enabled healthcare weakens *adoptability* (or end-user uptake) (Wong, Turner, and Yee 2008) and ultimately investor confidence. Greenhalgh et al. (2017) goes further to argue that when 'the value proposition

of the technology [is] unclear, in terms of a viable business venture for its developer [e.g. historical low CPM adoption rates] or in terms of a clear benefit for patients and an affordable real-world service model', the result is technology 'nonadoption and abandonment'. Our interviews with researchers reiterate this view that better research-practice collaborations based on trust and transparency encourage sustained use of computational models and private investment. This is particularly crucial for the successful clinical translation of computational models developed by small and medium developers or public research institutions. For both need to build robust investor or public-funder's confidence in the commercial viability (end-user uptake) of their innovation to attract the substantial capital resources and regulatory expertise necessary to fund and drive the clinical evidence generation, evaluation, verification, and validation processes needed to reach the market. At the same time, there is a need for stakeholders to acknowledge that widespread adoption of computational models should only be envisaged with particular attention to the risks of algorithmic bias – whereby AI applications normed on certain patient populations (e.g. whose data are easily available and widely used) may not be readily usable (without considerable adaptation, if at all) for diverse populations (see, e.g. Miner et al. 2020; Schwartz and Blankenship 2014).

4. Conclusion

In this paper, we have shown that the road from bench to bedside for computational brain models needs to remain grounded in building better researcher-clinician trust relationships based on meaningful upstream collaboration that integrate model transparency and tacit knowledge. On the one hand, greater clinician trust contributes to the robustness of investor or public-funder's confidence in the viability (end-user uptake) of an innovation to reach and succeed in the market and increases the translation of computational brain models. On the other hand, the clinician, both as custodian of patient trust and welfare, and as the end-user of computational models, is uniquely placed to evaluate the patient benefit of allocating scarce time and capital resources to computational models instead of in other areas of patient care which may include low-tech investments such as employing more clinicians, updating aging equipment etc. Thus policies encouraging greater clinician engagement in deciding 'if a computational models is worth it?' will not only act as a check against faulty analytics but also lead to responsible adoption of computational models that do not merely encourage technology adoption for the sake of technological progress but for meaningful benefit to patient care.

Special note

Dr Saheli Datta Burton and Dr Tara Mahfoud contributed equally to the work and are joint first authors, with Dr Saheli Datta Burton as corresponding author.

Acknowledgements

We would like to convey our special thanks to Dr Edison Bicudo (University of Sussex, UK), Professor Alex Faulkner (University of Sussex, UK), and to all who participated in this research. The authors acknowledge support from the European Union's Horizon 2020 Research and Innovation Programme funding for the *Human Brain Project* (Special Grant Agreement 2 number 785907).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors acknowledge support from the European Union's Horizon 2020 Research and Innovation Framework Programme funding for the *Human Brain Project* (Special Grant Agreement 2 number 785907); and the United Kingdom Engineering and Physical Sciences Research Council grant for the PETRAS-2 project (grant number EP/S03562/1).

Notes on contributors

S. Datta Burton is a Research Fellow at the Department of Science, Technology, Engineering and Public Policy, University College London. She is the Principal Investigator for the Petras project Building Evidence for CoP Legislation (BECL), a Visiting Research Fellow at the Department of Politics, University of Vienna, and a member of the Ethics and Society subproject of the Human Brain Project from 2017 till 2020. Saheli is interested in the international political economy of emerging technologies with a focus on data-driven health and medicine.

T. Mahfoud is a Lecturer in the Department of Sociology at the University of Essex. Before that, she was a Research Associate in the Department of Global Health and Social Medicine at King's College London and a member of the Human Brain Project Foresight Lab.

C. Aicardi is senior research fellow in King's College London Foresight Laboratory, Department of Global Health & Social Medicine. Christine worked for many years in the Information and Communication Technologies industry before returning to higher education to pursue a PhD in Science and Technology Studies at University College London. She joined the Ethics and Society subproject of the Human Brain Project in 2014.

N. Rose is Professor of Sociology at King's College London, and Co-Director of King's ESRC Centre for Society and Mental Health. He was the Principal Investigator and Director of the Foresight Lab at King's for the Human Brain Project and a member of the Steering Committee of the Ethics and Society Division of the HBP from its inception until 2020.

ORCID

S. Datta Burton  <http://orcid.org/0000-0001-8268-9013>

C. Aicardi  <http://orcid.org/0000-0003-1112-7720>

References

- Allegaert, Karel, Anne Smits, and N. van den Anker Johannes. 2012. "Physiologically Based Pharmacokinetic Modeling in Pediatric Drug Development: A Clinician's Request for a More Integrated Approach." *Journal of Biomedicine and Biotechnology*, doi:10.1155/2012/103763.
- Bhugra, Dinesh, Allan Tasman, Soumitra Pathare, Stefan Priebe, Shubulade Smith, John Torous, Melissa R. Arbuckle, et al. 2017. "The WPA Lancet Psychiatry Commission on the Future of Psychiatry." *The Lancet. Psychiatry* 4 (10): 775–818. doi:10.1016/S2215-0366(17)30333-4.
- Caruana, Rich, Scott Lundberg, Marco Tulio Ribeiro, Harsha Nori, and Samuel Jenkins. 2020. "Intelligible and Explainable Machine Learning: Best Practices and Practical Challenges." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (August): 3511–3512.
- Datta, Saheli. 2018. "Emerging Dynamics of Evidence and Trust in Online User-to-User Engagement: The Case of 'Unproven' Stem Cell Therapies." *Critical Public Health* 28 (3): 352–362. doi:10.1080/09581596.2018.1446509.
- De Fauw, Jeffrey, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, et al. 2018. "Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease." *Nature Medicine* 24 (9): 1342–1350. doi:10.1038/s41591-018-0107-6.
- Doi, Kunio. 2007. "Computer-aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential." *Computerized Medical Imaging and Graphics* 31 (4): 198–211. doi:10.1016/j.compmedimag.2007.02.002.
- Doraiswamy, P. Murali, Charlotte Blease, and Kaylee Bodner. 2020. "Artificial Intelligence and the Future of Psychiatry: Insights from a Global Physician Survey." *Artificial Intelligence in Medicine* 102: 101753.
- Engel, Heiberg P.J. 2008. "Tacit Knowledge and Visual Expertise in Medical Diagnostic Reasoning: Implications for Medical Education." *Medical Teacher* 30 (7): e184–e188. doi:10.1080/01421590802144260.
- Epstein, Steven. 1996. *Impure Science: AIDS, Activism, and the Politics of Knowledge*. (Vol. 7). Berkeley: University of California Press.
- Fiscella, Kevin, Sean Meldrum, Peter Franks, Cleveland G. Shields, Paul Duberstein, Susan H. McDaniel, and Ronald M. Epstein. 2004. "Patient Trust: Is It Related to Patient-Centered Behavior of Primary Care Physicians?" *Medical Care*, 1049–1055. doi:10.1097/00005650-200411000-00003.
- Fox, Susannah, and Maeve Duggan. 2013. "Health online 2013." *Pew Research Center*. <http://www.pewinternet.org/2013/01/15/health-online-2013/>.
- Future Advocacy. 2018. "Ethical, Social and Political Challenges of Artificial Intelligence in Health."
- Garland, Ann F., Dena Plemmons, and Leita Koontz. 2006. "Research-practice Partnership in Mental Health: Lessons from Participants." *Admin. and Policy in Mental Health and Mental Health Services Research* 33 (5): 517–528. DOI: 10.1007/s10488-006-0062-2.
- Gheiratmand, Mina, Irina Rish, Guillermo A. Cecchi, Matthew RG Brown, Russell Greiner, Pablo I. Polosecki, Pouya Bashivan, Andrew J. Greenshaw, Rajamannar Ramasubbu, and Serdar M. Dursun. 2017. "Learning Stable and Predictive Network-Based Patterns of Schizophrenia and its Clinical Symptoms." *NPJ Schizophrenia* 3 (22): 1–12. doi:10.1038/s41537-017-0022-8
- Goldstein, Kurt. 1995 [1939]. *The Organism: A Holistic Approach to Biology Derived from Pathological Data in man*. Brooklyn, NY: Zone Books.

- Gray, Bradford H. 1997. "Trust and Trustworthy Care in the Managed Care Era." *Health Affairs* 16 (1): 34–49. doi:10.1377/hlthaff.16.1.34.
- Greenhalgh, Trisha, Joseph Wherton, Chrysanthi Papoutsis, Jennifer Lynch, Gemma Hughes, Susan Hinder, Nick Fahy, Rob Procter, and Sara Shaw. 2017. "Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-up, Spread, and Sustainability of Health and Care Technologies." *Journal of Medical Internet Research* 19 (11): e367. DOI: 10.2196/jmir.8775.
- Gusnard, Debra, A. Marcus, and E. Raichle. 2001. "Searching for a Baseline: Functional Imaging and the Resting Human Brain." *Nature Neuroscience Rev* 2: 685–694. . doi:10.1038/35094500
- Hall, Mark A., Fabian Camacho, Elizabeth Dugan, and Rajesh Balkrishnan. 2002. "Trust in the Medical Profession: Conceptual and Measurement Issues." *Health Services Research* 37 (5): 1419–1439.
- Holzinger, Andreas, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. "What Do We Need to Build Explainable AI Systems for the Medical Domain?" *arXiv preprint arXiv:1712.09923* 1–28.
- Ingebrigtsen, Tor, Andrew Georgiou, Robyn Clay-Williams, Farah Magrabi, Antonia Hordern, Mirela Prgomet, Julie Li, Johanna Westbrook, and Jeffrey Braithwaite. 2014. "The Impact of Clinical Leadership on Health Information Technology Adoption: Systematic Review." *International Journal of Medical Informatics* 83 (6): 393–405.
- Joshi, Indra, and Jessical Morley. 2019. "Artificial Intelligence: How to Get It Right. Putting Policy Into Practice for Safe Data-Driven Innovation in Health and Care. London: NHSX.
- Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
- Khushf, George. A. 2013. "Framework for Understanding Medical Epistemologies." *Journal of Medicine and Philosophy* 38: 461–486. doi:10.1093/jmp/jht044.
- Klein, E., Andrew J. Solomon, John R. Corboy, and James Bernat. 2016. "Physician Compensation for Industry-Sponsored Clinical Trials in Multiple Sclerosis Influences Patient Trust." *Multiple Sclerosis and Related Disorders* 8: 4–8. doi:10.1016/j.msard.2016.04.001.
- Larson, Eric B. 2013. "Building Trust in the Power of "big Data" Research to Serve the Public Good." *Jama* 309 (23): 2443–2444. doi:10.1001/jama.2013.5914.
- Leblond, Richard F. 2013. "An Epistemology for Clinical Medicine: An Argument for Reflection on the Ends of Medical Practice and Ways of Knowing with Implications for the Selection and Training of Physician." *Transactions of the American Clinical and Climatological Association* 124: 238–249. PMCID: PMC3715941.
- Liberati, Elisa G., Francesca Ruggiero, Laura Galuppo, Mara Gorli, Marien González-Lorenzo, Marco Maraldi, Pietro Ruggieri, et al. 2017. "What Hinders the Uptake of Computerized Decision Support Systems in Hospitals? A Qualitative Study and Framework for Implementation." *Implementation Science* 12 (1): 113.
- Logothetis, Nikos. K. 2008. "What We Can do and What We Cannot Do with fMRI." *Nature* 453 (7197): 869–878. doi:10.1038/nature06976.
- Luo, Jake, Min Wu, Deepika Gopukumar, and Yiqing Zhao. 2016. "Big Data Application in Biomedical Research and Health Care: A Literature Review." *Biomedical Informatics Insights* 8: BII–S31559.
- Mahfoud, Tara. 2014. "Extending the Mind: A Review of Ethnographies of Neuroscience Practice." *Frontiers in Human Neuroscience* 8 (359): 1–9.
- Malterud, Kirsti. 1995. "The Legitimacy of Clinical Knowledge: Towards a Medical Epistemology Embracing the art of Medicine." *Theoretical Medicine* 16: 183–198. doi:10.1007/BF00998544.

- Mcknight, Harrison D., and Norman L. Chervany. 1996. "The Meanings of Trust." Technical Report 94-04, Carlson School of Management, University of Minnesota. http://misrc.umn.edu/workingpapers/fullpapers/1996/9604_040100.pdf.
- Miner, Adam S., Albert Haque, Jason A. Fries, Scott L. Fleming, Denise E. Wilfley, G. Terence Wilson, Arnold Milstein, et al. 2020. "Assessing the Accuracy of Automatic Speech Recognition for Psychotherapy." *NPJ Digital Medicine* 3 (1): 1–8. doi:10.1038/s41746-020-0285-8.
- Miner, Adam S., Nigam Shah, Kim D. Bullock, Bruce A. Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. "Key Considerations for Incorporating Conversational AI in Psychotherapy." *Frontiers in Psychiatry* 10: 746. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813224/>.
- Mittelstadt, Brent. D, and Luciano Floridi. 2016. "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts." *Science and Engineering Ethics* 22 (2): 303–341. doi:10.1007/s11948-015-9652-2.
- Morley, Jessica, Caio CV Machado, Christopher Burr, Josh Cows, Indra Joshi, Mariarosaria Taddeo, and Luciano Floridi. 2020. "The Ethics of AI in Health Care: A Mapping Review." *Social Science & Medicine* 2020: 113172. doi:10.1016/j.socscimed.2020.113172.
- Murray, Robin M. 2017. "Mistakes I Have Made in My Research Career." *Schizophrenia Bulletin* 43 (2): 253–256. doi:10.1093/schbul/sbw165.
- Nature Biotechnology. 2012. "What Happened to Personalized Medicine?" *Nature Biotechnology* 30 (1), doi:10.1038/nbt.2096.
- Nuffield Council on Bioethics. 2018. "Artificial Intelligence (AI) in Healthcare and Research. Bioethics Briefing Note." *Nuffield Council on Bioethics*. May. <http://nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research.pdf>.
- Patel, Vikram, Shekhar Saxena, Crick Lund, Graham Thornicroft, Florence Baingana, Paul Bolton, Dan Chisholm, et al. 2018. "The Lancet Commission on Global Mental Health and Sustainable Development." *Lancet* 392 (10157): 1553–1598. doi:10.1016/S0140-6736(18)31612-X.
- Polanyi, M. 2009. *The Tacit Dimension*. Chicago & London: University of Chicago Press.
- PricewaterhouseCoopers. 2012. "Social Media 'Likes' Healthcare." PWC Health Research Institute." <https://www.pwc.com/us/en/health-industries/healthresearch-institute/publications/pdf/health-care-social-media-report.pdf>.
- Rockoff, J. D. 2016. "J & J to Stop Selling Automated Sedation System Sedasys." *Wall Street Journal*, 14 March. <https://www.wsj.com/articles/j-j-to-stop-selling-automated-sedation-system-sedasy-1457989723>.
- Roepstorff, Andreas. 2007. "Navigating the Brainscape: When Knowing Becomes Seeing." In *Skilled Vision*, edited by C. Grasseni, 191–206. Berghahn: Oxford.
- Rose, Nikolas. 2013. "What is Diagnosis For?" *Talk given at the Institute of Psychiatry Conference on DSM-5 and the Future of Diagnosis*. 4 June.
- Rose, Nikolas, and Caitjan Gainty. 2019. "Neurovisions." *Advances in Clinical Neuroscience and Rehabilitation* 18 (2): 17–18. <https://www.acnr.co.uk/wp-content/uploads/2019/01/ACNR-NDJ19-low-rez-17-18.pdf>.
- Saunders, Barry, F. 2008. *CT Suite: The Work of Diagnosis in the Age of Noninvasive Cutting*. Durham & London: Duke University Press.
- Schwartz, Robert C., and David M. Blankenship. 2014. "Racial Disparities in Psychotic Disorder Diagnosis: A Review of Empirical Literature." *World Journal of Psychiatry* 4 (4): 133. doi:10.5498/wjp.v4.i4.133.
- Shah, Nilay D., Ewout W. Steyerberg, and David M. Kent. 2018. "Big Data and Predictive Analytics: Recalibrating Expectations." *Jama* 320 (1): 27–28. doi:10.1001/jama.2018.5602.

- Simonite, T. 2016. "Automated Anesthesiologist Suffers a Painful Defeat." *MIT Technology Review*, <https://www.technologyreview.com/s/601141/automated-anesthesiologist-suffer-s-a-painful-defeat/>.
- Smith, Rachel C, Claus Bossen, and Anne Marie Kanstrup. 2017. "Participatory Design in an era of Participation." *International Journal of CoCreation in Design and the Arts* 13 (2): 65–69. doi:10.1080/15710882.2017.1310466.
- Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, et al. 2016. "Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel." *Stanford University, Stanford, CA*. <http://ai100.stanford.edu/2016-report>.
- Sullivan, Greer, Naihua Duan, Snigdha Mukherjee, Joann Kirchner, Dana Perry, and Kathy Henderson. 2005. "The Role of Services Researchers in Facilitating Intervention Research." *Psychiatric Services* 56 (5): 537–542. doi:10.1176/appi.ps.56.5.537.
- Thom, David. 2001. "Physician behaviors that predict patient trust. Stanford Trust SP." *Journal of Family Practice* 50, 323–328. <https://go.gale.com/ps/anonymouse?id=GALE%7CA74292253&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=00943509&p=AONE&sw=w>.
- Thorpe, Alistair, Miroslav Sirota, Marie Juanchich, and Sheina Orbell. 2020. "'Always Take Your Doctor's Advice': Does Trust Moderate the Effect of Information on Inappropriate Antibiotic Prescribing Expectations?" *British Journal of Health Psychology* 25 (2): 358–376. doi:10.1111/bjhp.12411.
- Von Schomberg, Rene. 2013. "A Vision of Responsible Research and Innovation." In *Responsible Innovation Ch 3*, edited by R. Owen, M. Heintz, and J. Bessant, 51–74. London: John Wiley.
- Watson, David S., Jenny Krutzinna, Ian N. Bruce, Christopher EM Griffiths, Iain B. McInnes, Michael R. Barnes, and Luciano Floridi. 2019. "Clinical Applications of Machine Learning Algorithms: Beyond the Black box." *BMJ* 364 (l886): 1–9. doi:10.1136/bmj.l886.
- Wessler, Benjamin S., Jessica Paulus, Christine M. Lundquist, Muhammad Ajlan, Zuhair Natto, William A. Janes, Nitin Jethmalani, Gowri Raman, Jennifer S. Lutz, and David M. Kent. 2017. "Tufts PACE Clinical Predictive Model Registry: Update 1990 Through 2015." *Diagnostic and Prognostic Research* 1 (1): 20. doi:10.1186/s41512-017-0021-2.
- Wong, Ming C., Paul Turner, and Kwang C Yee. 2008. "Involving Clinicians in the Development of an Electronic Clinical Handover System-Thinking Systems not Just Technology." *Studies in Health Technology and Informatics* 136: 490–495. PMID: 18487779.
- Wynne, Brian. 2006. "Public Engagement as a Means of Restoring Public Trust in Science – Hitting the Notes, but Missing the Music?" *Community Genetics* 9 (3): 211–220. doi:10.1159/000092659.



Truth from the machine: artificial intelligence and the materialization of identity

Os Keyes^a, Zoë Hitzig^b and Mwenza Blell^c

^aDepartment of Human Centered Design & Engineering, University of Washington, Seattle, WA, USA;

^bEdmond J. Safra Center for Ethics, Harvard University, Cambridge, MA, USA; ^cPolicy, Ethics and Life Sciences Research Centre, School of Geography, Politics and Sociology, Newcastle University, Newcastle-upon-Tyne, UK

ABSTRACT

Critics now articulate their worries about the technologies, social practices and mythologies that comprise Artificial Intelligence (AI) in many domains. In this paper, we investigate the intersection of two domains of criticism: identity and scientific knowledge. On one hand, critics of AI in public policy emphasise its potential to discriminate on the basis of identity. On the other hand, critics of AI in scientific realms worry about how it may reorient or disorient research practices and the progression of scientific inquiry. We link the two sets of concerns—around identity and around knowledge—through a series of case studies. In our case studies, about autism and homosexuality, AI figures as part of scientific attempts to find, and fix, forms of identity. Our case studies are instructive: they show that when AI is deployed in scientific research about identity and personality, it can naturalise and reinforce biases. The identity-based and epistemic concerns about AI are not distinct. When AI is seen as a source of truth and scientific knowledge, it may lend public legitimacy to harmful ideas about identity.

KEYWORDS

Artificial intelligence; materialization; epistemology; reification; naturalization; race; sexuality; disability

1. Introduction

What happens when scientists go *looking* for identity with algorithmic tools? How is AI deployed to contextualize and justify their practices and results? And what can inquiring into this tell us about how we might critique and understand AI in more nuanced ways?

To answer these questions, we examine two distinct case studies – first, the foregrounding of machine learning systems in searches for the hypothesized genetic origins of autism, and second, the use of computer vision systems to trace the appearance of sexuality in facial structures. Two particular lines of concern addressed in this article are the impacts AI might have on how we learn and ‘know’ things, particularly in scientific research, and the ways in

which AI might reinforce or reshape ideas of identity in wider society. This article explores how these dual sets of concerns converge in peculiar ways. These case studies illustrate how such uses of AI rely upon and further reinforce a notion of identity as fixed, natural and essential. We argue that the cultural authority given to AI, combined with its increasing use in science and policy, creates the possibility of revitalizing and re-entrenching notions of identity and difference as ‘true,’ ‘objective’ and ‘real.’

While this materialization of difference is not new to AI, the increasingly ubiquitous use of these technologies makes it essential that researchers into the scientific and social impacts of AI attend to the sites and cultures in which such materialization occurs. We argue this attentiveness requires both contextual and historical awareness because the social impacts, whether positive or negative, are shaped by the spaces in which AI is deployed.

Inquiring into the impact of AI requires that we first confront what, precisely, AI is. As Krafft et al. note, AI suffers from a ‘definitional disconnect’ (Krafft et al. 2019): many people disagree over what the term means, even within the same discipline. Even after coming up with a formal definition, systems can be thought about ‘technically, computationally, mathematically, politically, culturally, economically, contextually, materially, philosophically, ethically, and so on’ (Kitchin 2017).

We approach the topic from a Science and Technology Studies (STS) perspective: that is, we treat technologies as evolving, not ‘in a vacuum,’ but ‘in the social world, being shaped *by* it and simultaneously *shaping* it’ (Law 2004). Therefore, here we define AI and examine it as constituting:

- (1) *AI as a technology*: the machine learning systems themselves, accompanied by the ‘big data’ they depend on (Gillespie 2014). The shape, affordances, and constraints of AI technologies influence how they are seen, engineered, and deployed.
- (2) *AI as a social practice*: the work of building, deploying and articulating AI. As Nick Seaver puts it, ‘social structures emboss themselves onto digital substrates; software is a kind of print left by inky institutions’ (Seaver 2018). AI cannot be understood independently of the individuals, collectives and institutions that use and shape it.
- (3) *AI as a set of mythologies*: the rhetoric and cultural narratives that define and modulate perceptions of AI, from scientific communities to popular culture. These strongly overlap with notions of ‘imaginaries,’ narratives which ‘describe attainable futures and prescribe the images of futures that should be attained’ (Felt et al. 2016, 754) but are not always future-oriented: they frequently instead address perceptions of what AI can do *right now*. Such mythologies ‘condition not only the perception of technology within the public’ but also ‘the professional culture of those who have

produced the technical innovations and helped their development' (Natale and Ballatore 2017, quoting Ortoleva 2009)

This three-part frame reveals considerable discontent with AI. It has been argued, for instance, that by altering public notions of truth and reshaping what it means to be a subject in society, AI undermines democracy (Helbing et al. 2019; Zimmer et al. 2019; Stark 2018). Going further, AI systems and the data that underlie them are now frequently dramatized as fundamentally rewriting the fabric of society – and 'human nature' along with it (Maclure 2019), for better or worse. Rather than explore the implications of AI as a technology, practice, and set of mythologies for society generally; however, in this article, we use this three-part frame to examine the interplay of AI with *scientific practices, ideas of knowledge, and social identities*. After historicising and contextualizing discontent surrounding these issue, our analysis explicates the consequences of applying *algorithmic* scientific practices to questions of identity.

1.1. AI and scientific knowledge

Paralleling the transformative impact it is widely predicted to have on society generally, advocates of AI often portray it as fundamentally shifting how science is conducted, as well as the way knowledge is generated and validated. This includes not only the traditional sciences, but social inquiry as well, with a growing community of researchers claiming to practice a 'computational social science' that 'leverages the capacity to collect and analyse data with an unprecedented breadth and depth and scale' (Lazer et al. 2009).

Such claims raise concerns related to the technical capabilities and directions of AI. Critics regularly point out that many AI systems are 'black boxes' – it is often difficult if not impossible to of unpack *why* an AI model produced a particular output. In practice, the conclusions of these systems resist human interpretation 'even for those with specialized training, even for computer scientists' (Burrell 2016; Ananny and Crawford 2018). For McQuillan (2017), this difficulty raises the possibility of what he calls 'machinic neoplatonism', a world in which scientists approach algorithms as revealing pieces of some fundamental, universal truth. This approach results not in a free spirit of scientific inquiry, but a ritualized system that treats algorithms as transcendent oracles. In such cases, AI becomes less a revolution than a *regression*, constraining the depth of scientific understanding to a superficial and outdated form of positivism.

The possible negative consequences of this regression are magnified by a second set of technical concerns: the epistemic downsides of 'Big Data.' At the centre of AI's constellation of promises lies the idea that data collection at unprecedented scales and levels of detail – in combination with clever algorithms for computing it – will make visible correlations and connections that

were hidden from merely human eyes. Rather than painstakingly (which is to say, *manually*) exploring particle interactions or protein embeddings, a computational model can rapidly simulate all possibilities and relations (Carrasquilla and Melko 2017; Yang et al. 2018). Yet critics contend that because such large datasets ‘have to contain arbitrary correlations,’ the returns to scale diminish into yet another regression, for ‘too much information tends to behave like very little information’ (Calude and Longo 2017). Consequently, algorithmically revealed correlations way may not actually be *accurate* (L’heureux et al. 2017). Further, a dependence on Big Data simultaneously occludes questions where the available data is *not* ‘Big’ enough, leading to fundamental changes not only in scientific methods, but in the very questions science tries to answer (Bowker 2014).

That these technical promises may not be met does not preclude them from altering and distorting concepts of scientific and ordinary knowledge. Cultural mythologies and imaginaries about ‘what data can do’ reshape work, practices and values even if their promises are not (and may never be) kept. Given what Dumit refers to as the ‘taboo nature of subjectivity in science’ where ‘every possibility of subjectivity must be eliminated in order to produce something reliable – that is, something real, something known,’ it is unsurprising that within science, the mythologies of AI possess considerable power. Scientific researchers often valorize AI, in practice and in public, as the pinnacle of ‘automation, which stands as the opposite of interactivity’ (Dumit 2004, 122). In other words, AI holds the potential to become the apotheosis of scientific objectivity. This constitutes an ‘epistemological hazard’ (Elish and Boyd 2018, 58), since the potential objectivity of AI often amounts to nothing more than a veneer of certainty, and moreover, creates alarming interpretive dynamics. For example, even if research processes using AI do not produce more ‘accurate’ results, these processes may nonetheless be interpreted as grounded in additional certainty – simply by deploying the rhetoric of AI. The risk is not the power of algorithms *per se*, but ‘the power of the notion of the algorithm... the way that notions of the algorithm are evoked as a part of broader rationalities and ways of seeing the world... envisioned to promote certain values and forms of calculative objectivity’ (Beer 2017).

1.2. AI and identity

A distinct set of concerns have been raised about what the deployment of and dependency on widespread algorithmic systems might mean for questions of *identity*. In a world increasingly filled with automated classificatory systems based on algorithmic inference – in everything from advertising and Internet searching to medicine and legal decisions – the power of such systems is tremendous. It is hard to imagine how these systems could avoid affecting

individuals' sense of identity or producing differential effects on the *grounds* of identity.

Some immediate issues concern biased or discriminatory outcomes. Both theorists and practitioners have demonstrated ways in which algorithmic systems produce disparate impacts for different populations – impacts which frequently negatively effect trans people, people of colour, and/or the poor (Eubanks 2018; Noble 2018; Keyes 2018). While 'biased data' is often identified as the cause, some researchers caution that the answer is more complex. Even with seemingly neutral data, designers and users ultimately carry particular notions of identity with them into the lifeworlds where systems are built and deployed, layering their own expectations of what gender, race or class mean onto and into AI (Van der Ploeg 2012; M'charek, Schramm, and Skinner 2014). Further, given that the premise of many systems is one of *classification*, there are questions about whether the fluidity and malleability of identity can be adequately represented at all (Keyes 2019). Thoughtful scholars of gender, post-colonial studies and critical race theory have extensively documented the long histories of colonialism, violence, and oppression that come with efforts to restrict something as flexible as the self to fixed and measurable forms (Hames-Garcia 2011; Lugones 2016; Bhagat 2006; Thompson 2015).

At a more conceptual and existential level, concerns have been raised about the ways in which AI might *rewrite* ideas of identity altogether. Echoing the concerns of Katja De Vries (2010), John Cheney-Lippold summarizes such worries in hypothesizing a 'a new analytical axis of power: the digital construction of categories of identity' (Cheney-Lippold 2011, 172). What race, gender or other aspects of identity 'mean' – their consequences, how they are assessed, how those placed in different categories understand themselves, and how accessible these meanings even are to them – are altered. This reinforcement of a notion of identity as an external quality that can be 'objectively' inferred produces a new form of control 'which works not just on the body nor just on the population, but in how we define ourselves and others' (Ibid., 177).

1.3. Contextualizing discontent

Reflecting on the work summarized above, we see several interesting absences. One is that, as mentioned above, many critics echo AI advocates in describing the relationship between AI and society as deterministic. Both implicitly treat AI as fundamentally new, without history. But there are good reasons to be sceptical of such a view, both for understanding AI in science and its consequences for identity. From an STS perspective, many of the hypothesized consequences of AI are neither specific to AI nor particularly new. The 'black-boxing' of decisions in science is understood as a longstanding practice that is more or less inevitable: for scientific ideas to be regarded as *certain* and worthy of adoption, black boxing is precisely what is required to avoid a

situation where each researcher in a scientific network must comprehend the complete depths of every part of it before making any movement through it (Latour 1987). Similarly, technological artefacts have long mediated the relationship between scientists and their objects of study. Within science, technological mediation is typically regarded as an epistemological necessity, constituting the core of ‘mechanical objectivity’, in which scientific rigour is strongly associated with the degree to which the scientist is removed from the process of knowledge creation (Galison and Daston 2007).

As in scientific practice, so, too, in identity. From genetics to online dating platforms to lung measurements, researchers have regularly highlighted the long history of seemingly innocuous technoscientific systems being premised on and reinforcing of racial disparities and gender stereotypes, as well as ideas of sexuality and disability (Roberts 2011; Bivens and Haimson 2016; Braun 2014). This is not only a matter of creating or perpetuating inequalities but also of altering notions of what identity is and where it is to be found. The creation of new technologies of measurement has always led to – and in many cases been driven by – the opportunity to change conceptions of *what is being measured*. A prominent case study would be the ‘penile plethysmograph’, an instrument to measure arousal in phalluses: designed and adopted for inquiries into sexuality, the idea was to test the hypothesis that ‘for men, arousal *is* orientation’ (Waidzunus and Epstein 2015). To the extent that it cannot be extricated from culture, identity is and has always been mediated by technology. Interrogations of technologies that ignore this fact risk hiding the wider mechanisms of power and knowledge that enabled the technology’s adaption in the first place (Keyes 2018).

We recognize the historical roots of these phenomena as contingent and contextual. STS has long emphasized that the ways in which people make sense of, use, and are impacted by technologies are themselves mediated by local circumstances: ‘the shapes of knowledge are always ineluctably local, indivisible from their instruments and their encasements’ (Geertz 1973, 4). As Karin Knorr Cetina (2009) has documented, scientific cultures, and their approaches to technologies or knowledge, vary widely, between fields and individual laboratories. As a result, we lose something if we treat the likely impacts of AI as uniform and predetermined. We instead need to carefully examine how AI, whether in technological or mythological form, is situated and used. As Seaver poetically puts it, we must ‘examine the logic that guides the hands, picking certain algorithms rather than others, choosing particular representations of data, and translating ideas into code’ (Seaver 2019, 419). Similar concerns have been raised by Taina Bucher, who takes issue with the practice of treating AI as a determined and deterministic thing. While their focus is on the process of *developing* algorithms, the very fact that algorithms exist in a wider assemblage of people and datasets and conclusions means that this applies just as well to questions of algorithms’ *use* (Seaver 2019; Bucher 2016). New technologies and the

circumstances of their adoption work ‘with, across and through [existing] conventions, technologies and communities’ (Coopmans et al. 2014, 5). While ethnographic work examining the use of data science in commercial contexts regularly acknowledge the importance of historical contingency and social context, the same is not true of the research into how AI is used in *science* (Passi and Sengers 2020; Wolf and Paine 2018).

We believe, then, that work seeking to explore the uses of AI systems in and their consequences for science must be both contextualized and historicised. We are, of course, hardly the first people to make such a claim. Although her work is largely focused on surveillance and racialized violence, Simone Browne makes a similar argument in calling for a ‘critical biometric consciousness’ that understands technologies such as facial recognition as the latest evolution in a long line of similar mechanisms (Browne 2015). And as Browne’s work (and our reference to it) suggests, we believe that efforts to historicise and contextualize technologies of identity and of scientific epistemology will find that these two areas of concern are not distinct.

This leads us to the final – and most concerning – absence in the work we have examined. Concerns about the shaping of scientific knowledge and the shaping of identity under AI are rarely in conversation with each other. There are some exceptions that look at how algorithmic tools for measuring identity within science change what it means to *know* identity (Keyes 2018), but overall, there is little interplay. This is concerning precisely because of history, and because of context. Scientific knowledge has played a central role in how we understand identity for centuries, and vice versa (Downing, Morland, and Sullivan 2015; Samuels 2014).

In sum, then, we aim to historicise and contextualize concerns about the impact of AI on scientific practice, about the effect of AI on notions of identity, and about the interplay between these two types of impacts. Our question is: What happens when the identity-based and scientific uses of AI intersect? What happens when AI is deployed by scientists to ‘find’ identity? What social worlds are produced, what ideas are reinforced, and what are the dangers that result? And how do local histories and contexts play a role in determining the answers?

2. AI, science and identity

To answer these questions we rely on two case studies – one, the deployment of AI as a tool for research into the aetiology of autism, and the other, the deployment of AI to find evidence of sexuality or race-based differences in facial structure. In both cases, we explore what happens when AI is used in scientific research *into* identity, the historical and contextual factors in its adoption, as well as the legitimization and consequences of the results. Rather than simply aiding or obscuring scientific inquiry, we argue that AI serves not to find the

‘truth’ of identity but to naturalize a particular view of it – one that, unsurprisingly, conforms with status quo assumptions.

This process of naturalization aligns with what Campolo & Crawford term ‘enchanted determinism’:

a discourse that presents [AI] as magical, outside the scope of present scientific knowledge, yet also deterministic, in that [AI] can nonetheless detect patterns that give unprecedented access to people’s identities, emotions and social character (Campolo and Crawford 2020, 1)

Precisely because AI is so often treated as capable of revealing otherwise unknowable (and therefore *unquestionable*) truths, this has worrying implications. But as we hope to demonstrate, where and how ‘enchanted determinism’ appears, as well as the implications that follow from it, are informed by the contexts and histories surrounding the scientific use of AI.

2.1. Autism, algorithms and genetics

One site of scientific work that has seen widespread adoption of AI is *autism research*. Seeking to find genetic ‘origins’ and neurological indicators of autism, researchers have increasingly turned to machine learning techniques to grapple with their data (Sato et al. 2013; Kassraian-Fard et al. 2016). AI is seen as capable of providing a path through the uncertainty and heterogeneity that characterizes current research. By drawing on larger datasets and methodologies for finding ‘global, complex and potentially multimodal patterns of abnormalities that cannot be efficiently identified with univariate methods’ (Ecker 2011), researchers hope to winnow some consistent signal from the noise.

When we talk about uncertainty and heterogeneity, however, what we are ultimately talking about is scientific failure. The goal is to find the biological *origins* of autism: ideally, a single, consistent marker. Yet this smoking gun consistently eludes scientific inquiry. To the researchers, this is a problem stemming from technical complexity; genomes, brains and human development are extraordinarily complex, but people and technologies are limited in their ability to grasp such complexity. Perhaps there are multiple biological sources or multiple pathways of development that result in autism (Fitzgerald 2017).

Another explanation, however, is that what researchers are looking for is not a fixed, natural ‘kind,’ or some immutable state of being. Rather, the nature of ‘autism’ may be a moving target, one shaped by changes in behaviour, social norms and diagnostic criterion, as suggested by a range of work in the social sciences. Gil Eyal and collaborators have traced, for example, the role that race played in initially stabilizing the diagnosis of ‘autism’ – and the ways in which changes to the diagnostic criterion were driven not simply by refined

researcher knowledge, but by wider cultural narratives and lobbying efforts around demedicalization (Eyal 2010). More generally, researchers have pointed to autism as a canonical example of what Ian Hacking refers to as ‘looping effects’ – where the experiences of classes of people (such as ‘autistic people’) interacting with infrastructures of treatment and meaning lead to changes in their behaviour and presentation, necessitating changes in the diagnostic criteria and infrastructures, producing changes in the population, and so on (Woods 2017). The result is a tangle of different meanings, and wildly divergent diagnostic criteria and cultural meanings at different points in time, to the extent that some researchers have simply concluded that ‘autism’s inherent heterogeneity lends it an ontological indeterminacy, meaning that exactly what autism is can never be known’ (Hayes et al. 2020, p.827).

Such perspectives have not stopped researchers from applying machine learning to this problem – thus treating it as a technical problem – and claiming some inherent truth to what their algorithms find.¹ As an illustrative example, we point to Zhou and colleagues’ ‘Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk’ (Zhou et al. 2019). As suggested by the title, these researchers built a deep learning system to analyse the genomes of individuals with autism, seeking to identify candidate genes and mutations in a more thorough and nuanced fashion than previously possible. Such an approach fits nicely with the epistemic imaginaries surrounding AI – machine learning systems are widely regarded as capable of picking complex patterns and signals out of data previously dismissed as noise.

This work (and others like it) does not confront the complexity and socially shaped nature of what ‘autism’ is, nor does media coverage of it. Zhou et al.’s paper was reported by their sponsoring institution as using AI to detect ‘a new class of mutations behind autism’, and promised that ‘this powerful method is generally applicable to discovering such genetic contributions to any disease’ (Schultz 2019). More independent headlines ran as ‘AI Discovers Causes of Autism in Uncharted DNA’ and ‘AI detects new class of genetic mutations behind autism’ (Genetic Engineering and Biotechnology News 2019; Tribune 2019), attributing certainty to the outcome of the study and fixed biological explanations of what ‘autism’ is. The study’s authors are quoted as saying that an AI-based approach ‘transforms the way we need to think about the possible causes of those diseases’, thus treating AI as a profound and fundamental shift in scientific research. More crucially, the language of ‘powerful methods’ and the emphasis on AI in particular implies that, to the authors – and the desired readers – AI should be understood to lend a stronger valence of truth to the study’s results than otherwise possible. When applied to studies such as this one, the rhetoric of AI as a powerful means of discovering

¹In fairness to these researchers, glossing over the complex history of autism as a concept is a practice autism researchers have long engaged in, frequently preferring (as most scientists prefer) a linear history of new, improved truths inexorably overtaking old falsehoods (Verhoeff 2013; Hollin 2014).

the truth further cements the value and authority of the results, as well as the researchers who produce them.

This, then, is a classic example of scientific work that naturalizes fixed ideas of identity and personality, one that uses the mythology of ‘AI’ to boost its legitimacy – precisely what Campolo & Crawford describe as enchanted determinism. In addition, however, it is a case study in how history and context can inform scientific work and its outcomes. AI was not adopted simply because it was *technically* suitable or because AI is being adopted everywhere; rather, it fits the pre-existing processes and approaches that autism researchers take in organizing and structuring their work. As Jennifer Singh has documented, autism research includes vast assemblages of researchers, state actors and (largely parent-driven) advocacy organizations. Within this assemblage, large sums of money are provided to researchers in the hunt for autism’s aetiology – a hunt that has, thus far, failed. The need to justify existing and new expenses has led, amongst other things, to a focus by both researchers and funding agencies on bleeding-edge technologies using the largest datasets and consortia of researchers available (Singh 2015). AI – a technology culturally understood as singularly suited to thorny and unsolvable problems, and as a quintessential ‘bleeding edge’ approach to research – is thus ideally suited to adoption.

Similarly, is not solely due to the use of AI that the results carry the ring of objective truth to their audience. What we are looking at here, after all, is AI *paired with* genomics and neuroscience, both of which enjoy extant and influential mythologies of truth. The idea of the ‘cerebral subject’ – the locating of the truth of identity and self-hood in the material structures of the brain – has become highly powerful in both research and wider society and is part of the shift in research towards genomic and neuroscientific explanations (Rose and Abi-Rached 2013; Ortega and Choudhury 2011). With autism, in particular, there is a strong neuroscientific and genomic bent within research as well as advocacy organizations, including the ‘neurodiverse’ movement of autistic individuals who have seized on signs of neurological differences to demand the treatment of autistic people as representing a distinct subculture and way of being, rather than a biological failure. AI’s power here is neither a result of ‘AI’ in the abstract, or ‘AI’ as a universal. Rather, it acts as a catalyst to pre-existing lines of research and ways of identifying truth.

2.2. Facing down sexuality

Shifting from the inner workings of the cell to the outer workings of the face, our second case study is the now infamous ‘gaydar’ study by two Stanford University researchers, Yilun Wang and Michael Kosinski. The reception of this study is indicative of a broader pattern in which harmful and previously discredited theories are reinvigorated by the appearance of an AI-based argument in their support.

While the idea of ‘detecting’ homosexuality in measurements of the body originated in the late nineteenth century, the history of research *disputing* the validity of any biological link is almost as long. Nevertheless, efforts to ‘find’ sexuality in the body continue, albeit with somewhat less academic prominence than in the past (Terry 1999). Importantly, whatever the formal status of this research in academia, biological theories of sexuality are prominent in wider culture and society and have played a powerful role in both grounding and undermining efforts to address homophobia. It is the conjoining of these theories with AI – and the power and prominence of each – that makes Wang & Kosinski’s study so potent. Even if these AI research findings or methodological approaches are at some point debunked by other members of some scientific communities, their study is both highly cited compared to other work on the same topic using other methods, and has had a great impact on public consciousness thanks to the wide publicity it received.²

In 2018, Wang & Kosinski publicly released the preprint of their (subsequently accepted) paper ‘Deep neural networks are more accurate than humans at detecting sexual orientation from facial images’, receiving substantial media coverage. Using a dataset of facial images acquired from an online dating website, the researchers attempted to create a machine learning system that would identify the sexuality of a photograph’s subject from analysing it. Such an approach, they argued, was well suited to be generalized, allowing researchers to ‘boost our understanding of the origins and nature of a broad range of psychological traits, preferences, and psychological processes’ (Wang and Kosinski 2018, 30-31).³

For Wang & Kosinski, machine learning serves not just to enable analysis to scale (as in the case of autism research) but to discern fundamental truths that are simply too subtle for the human mind. They argue that ‘people may lack the ability to detect or interpret [the differences]. It is possible that some of our intimate traits are prominently displayed on the face, even if others cannot perceive them’ (Wang and Kosinski 2018, 4-5). This is what motivates their choice of *machinic* vision: the belief that ‘The links between facial features and sexual orientation... may be stronger than what meets the *human* eye.’ (Wang and Kosinski 2018, our emphasis). Again, this framing is a quintessential example of enchanted determinism, one echoed in other studies designed to infer aspects of identity and personality through similar techniques (Calvo and D’Mello 2010).

But the reason Wang & Kosinski’s study was so controversial, and seen as so dangerous, is not simply a matter of AI systems naturalizing identity. Rather, it

²As of writing, the paper has received 305 citations in under two years, along with coverage in the *New York Times*, *The Guardian*, *The Economist* and the *Financial Times*.

³The authors later claimed their true motivation in writing and publishing this work was to demonstrate the *dangers* here. Why this required them to explain, in great detail, how to build a ‘gay face’ detector, is unclear. We should be profoundly grateful that they did not, for example, decide to contribute to nuclear disarmament, since they would presumably have done so by designing, building, and detonating a hydrogen bomb before publishing the schematics online – just to ensure everyone really understood the dangers.

has to do with the way their work resonated with problematic studies of the past. The sciences have a long history of inquiry into the ‘nature’ and roots of gender and sexuality, one that has (at various times and in various disciplines) encompassed psychiatry, neuroscience, genetics, and endocrinology (Terry 1999; Brookey 2002). Wang & Kosinski explicitly located their work within this history, arguing for the viability of studying differences between the faces of people of different sexual orientations by pointing to research that claims that

same-gender sexual orientation stems from the underexposure of male fetuses or overexposure of female fetuses to androgens that are responsible for sexual differentiation... gay men should tend to have more feminine facial features than heterosexual men, while lesbians should tend to have more masculine features than heterosexual women. Thus, gay men are predicted to have smaller jaws and chins, slimmer eyebrows, longer noses, and larger foreheads; the opposite should be true for lesbians. (Wang and Kosinski 2018, 6)

These hypotheses about sexuality – while presented as strongly grounded by the authors – might *generously* be described as ‘unsupported by any adequate data’ (Hird 2004; Jordan-Young 2011). But they are regularly treated as credible in both sexology and folk understandings of sexuality, with profoundly harmful consequences; historians and sociologists have extensively tracked the role of this sort of essentialism in debates over the societal legitimacy of queer lives, and the construction and legitimization of individual lives. At the macro scale, these hypotheses have been used to justify queerphobia, due to the implication that queer people are biologically aberrant, and to reinforce rigid ideas of gender and sexuality (Waidzunus 2015; Terry 1999). At the micro, while these hypotheses are at best inadequately evidenced, the same cannot be said of the extensive research demonstrating the ways in which internalized essentialism reinforces individual homophobia (Haslam and Levy 2006; Morandini et al. 2015).

The use of AI in this study further grants those theories public legitimacy – and although it was thoroughly critiqued by the broader scientific community, its results remain in the social imaginary, their uptake aided by the veneer of credibility that ‘AI’ provides. In other words, what makes Wang & Kosinski’s work so resonant, and so dangerous, is the wider context in which it was undertaken. The danger is not the singular contribution of AI, but the result of AI being used to reinforce and cement existing discourses of harm. The same is true of other efforts to infer identity from the face, in domains from disability and gender to race (Hashemi et al. 2012; Bautista et al. 2015; Fu, He, and Hou 2014). In each case, the power of algorithmic systems is not simply the credibility that AI’s mythology lends to the results, but the way that such results resonate with existing folk understandings of identity’s visibility and biological fixity.

Our argument here is not that this danger is novel. To the contrary, science has often been deployed or taken up in precisely this way. ‘AI’ is simply to the twenty-

first century what ‘genetics’ was to the twentieth, or anthropometrics to the nineteenth – a tool of inquiry that, buoyed by both popular and academic understandings of it as an unprecedented and unimpeachable source of truth, is deployed to legitimize (rhetorically or methodologically) the same old schemes of division and disparity. Our argument is simply that inquiries into the nature and risks of how AI is being (conceptually and methodologically) deployed around identity must begin from the recognition that no research exists in a vacuum. Existing understandings of the fixity of identity – and where that fixity is to be located – inform both the shape of studies and their uptake. Just as the adoption of AI in neuroscience depended in part on existing infrastructure and expectations, the *consequences* of Wang & Kozinski’s work do not come solely from the rhetorical power of ‘AI’. Rather, they come from the marriage of AI’s mythology with existing essentialist explanations, both folk and scientific: the way in which AI is used to demonstrate the validity of what ‘everyone already knows’.

3. Discussion and conclusion

Through our two case studies, we have simultaneously shown the risks of deploying AI as a technology in scientific inquiry into identity, and the way that these risks – and the viability of these deployments – are contingent not only on AI, but on the history and context of identity and/or science. This has several clear implications and paths forward for research into the nature of AI and its impacts.

First, and most generally, we wish to reiterate the need for researchers – both those who laud algorithmic systems, and those who critique – to attend to context and history. AI is neither an entirely novel plague nor a universal panacea; it is a technology (and mythology, and set of practices) that will appear in different forms to different observers in different spaces. Rather than buying into the rhetoric that ‘everything is different now’, we should instead ascertain what is different, in what ways, and under what circumstances. As this article demonstrates, while there is certainly novelty in AI, many of the harms emerging from its deployment are quite old.

Second, it is urgent that such inquiry move beyond treating various areas of deployment and concern as entirely distinct. Just as with AI’s mythology, technology, and practices, sites and types of research can rarely be easily parsed out into ‘epistemological’ versus ‘identity’, or ‘theoretical research’ versus ‘practice’. In our work, we have sought to demonstrate both that identity-based concerns about AI are not distinct from epistemic concerns, and that private industry cannot be understood as the exclusive or primary space where AI might be used and cause harm. In the case of the former, we see (over and over again) the mythology of AI being used to reinforce existing stereotypes of disability and sexuality, further legitimizing their violence. In the latter case, we highlight the ways in which our understandings of identity and knowledge – though

undoubtedly more shaped by the private sector under neoliberalism than was previously true – are still strongly tied to scientific ideas and work.

Our ultimate hope, then, is that this paper will be taken as a prompt for other researchers – both those inquiring into AI and those considering using it – to critically contextualize and historicize their sites of work, and to attend to the urgent need to consider the ways in which existing, rather than entirely novel, forms of violence continue to haunt technology's effects.

Acknowledgments

Our thanks go, first and foremost, to each other. We are additionally grateful to Nikki Stevens, Claire and Margaret Hopkins, Adam Hyland, and our anonymous reviewers and editors for their ongoing support.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially funded by a Microsoft Ada Lovelace Fellowship.

Notes on contributors

Os Keyes is a PhD candidate at the University of Washington, where they study the interplay of identity, infrastructure and (counter)power. They are a frequent media commentator and public scholar, and an inaugural winner of the Ada Lovelace Fellowship.

Mwenza Blell is currently a Rutherford Fellow aliated to Health Data Research UK, a Newcastle University Academic Track Fellow, and a Grant Researcher at Tampere University. A biosocial anthropologist, her research draws from ethnography to examine intransigent and often invisible structures of injustice.

ZoLe Hitzig is a PhD candidate in economics at Harvard University. Her graduate research has been supported by fellowships from the Edmond J. Safra Center for Ethics at Harvard, Microsoft Research and the Forethought Foundation.

References

- Ananny M., and K. Crawford. 2018. "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20 (3): 973–989.
- Bautista M. A., A. Hernández-Vela, S. Escalera, L. Igual, O. Pujol, J. Moya, V. Violant, and M. T. Anguera. 2015. "A Gesture Recognition System for Detecting Behavioral Patterns of Adhd." *IEEE Transactions on Cybernetics* 46 (1): 136–147.
- Beer D. G. 2017. "The Social Power of Algorithms." *Information, Communication and Society* 20: 1–13.

- Bhagat R. B. 2006. "Census and Caste Enumeration: British Legacy and Contemporary Practice in India." *Genus* 62: 119–134.
- Bivens R., and O. L. Haimson. 2016. "Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers." *Social Media+ Society* 2 (4). doi: doi:10.1177/2056305116672486.
- Bowker G. C. 2014. "Big Data, Big Questions— the Theory/Data Thing." *International Journal of Communication* 8: 1795–1799.
- Braun L. 2014. *Breathing Race into the Machine: The Surprising Career of the Spirometer From Plantation to Genetics*. Minneapolis: University of Minnesota Press.
- Brookey R. A. 2002. *Reinventing the Male Homosexual: The Rhetoric and Power of the Gay Gene*. Bloomington: Indiana University Press.
- Browne S. 2015. *Dark Matters: On the Surveillance of Blackness*. Durham: Duke University Press.
- Bucher T. 2016. "Neither Black nor Box: Ways of Knowing Algorithms." In: *Innovative Methods in Media and Communication Research*, 81–98. Cham: Springer.
- Burrell J. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1). doi: doi:10.1177/2053951715622512.
- Calude C. S., and G. Longo. 2017. "The Deluge of Spurious Correlations in Big Data." *Foundations of Science* 22 (3): 595–612.
- Calvo R. A., and S. D'Mello. 2010. "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications." *IEEE Transactions on Affective Computing* 1 (1): 18–37.
- Campolo A., and K. Crawford. 2020. "Enchanted Determinism: Power Without Responsibility in Artificial Intelligence." *Engaging Science, Technology, and Society* 6: 1–19.
- Carrasquilla J., and R. G. Melko. 2017. "Machine Learning Phases of Matter." *Nature Physics* 13 (5): 431.
- Cetina K. K. 2009. *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.
- Cheney-Lippold J. 2011. "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control." *Theory, Culture & Society* 28 (6): 164–181.
- Coopmans C., J. Vertesi, M. E. Lynch, and S. Woolgar. 2014. *Representation in Scientific Practice Revisited*. Cambridge, MA: MIT Press.
- De Vries K. 2010. "Identity, Profiling Algorithms and a World of Ambient Intelligence." *Ethics and Information Technology* 12 (1): 71–85.
- Downing L., I. Morland, and N. Sullivan. 2015. *Fuckology: Critical Essays on John Money's Diagnostic Concepts*. Chicago: University of Chicago Press.
- Dumit J. 2004. *Picturing Personhood: Brain Scans and Biomedical Identity*. Princeton: Princeton University Press.
- Ecker C. 2011. "Autism Biomarkers for More Efficacious Diagnosis." *Biomarkers in Medicine* 5 (2): 193–195.
- Elish M. C., and D. Boyd. 2018. "Situating Methods in the Magic of Big Data and AI." *Communication Monographs* 85 (1): 57–80.
- Genetic Engineering and Biotechnology New. 2019. AI Finds Autism-Causing Mutations in 'Junk' DNA. *Genetic Engineering and Biotechnology News*. <https://www.genengnews.com/news/ai-finds-autism-causing-mutations-in-junk-dna/>.
- Eubanks V. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.
- Eyal G. 2010. *The Autism Matrix*. Cambridge: Polity.
- Felt U., J. Igelsböck, A. Schikowitz, and T. Völker. 2016. "Transdisciplinary Sustainability Research in Practice: Between Imaginaries of Collective Experimentation and Entrenched Academic Value Orders." *Science, Technology, & Human Values* 41 (4): 732–761.

- Fitzgerald D. 2017. *Tracing Autism: Uncertainty, Ambiguity, and the Affective Labor of Neuroscience*. Seattle: University of Washington Press.
- Fu S., H. He, and Z. G. Hou. 2014. "Learning Race From Face: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36 (12): 2483–2509.
- Galison P., and L. Daston. 2007. Objectivity.
- Geertz C. 1973. *Local Knowledge: Further Essays in Interpretive Anthropology*. New York: Basic books.
- Gillespie T. 2014. "The Relevance of Algorithms." *Media Technologies: Essays on Communication, Materiality, and Society* 167: 167.
- Hames-Garcia M. R. 2011. *Identity Complex: Making the Case for Multiplicity*. Minneapolis: University of Minnesota Press.
- Hashemi J., T. V. Spina, M. Tepper, A. Esler, V. Morellas, N. Papanikolopoulos, and G. Sapiro. 2012. "A Computer Vision Approach for the Assessment of Autism-Related Behavioral Markers." In: *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. 1–7, IEEE.
- Haslam N., and S. R. Levy. 2006. "Essentialist Beliefs About Homosexuality: Structure and Implications for Prejudice." *Personality and Social Psychology Bulletin* 32 (4): 471–485.
- Hayes J., R. McCabe, T. Ford, and G. Russell. 2020. "Drawing a Line in the Sand: Affect and Testimony in Autism Assessment Teams in the UK." *Sociology of Health & Illness* 42: 825–843.
- Helbing D., B. S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, J. Van Den Hoven, R. V. Zicari, and A. Zwitter. 2019. "Will democracy survive big data and artificial intelligence?" In: *Towards Digital Enlightenment*, 73–98. Cham: Springer.
- Hird M. J. 2004. *Sex, Gender, and Science*. Basingstoke: Palgrave Macmillan.
- Hollin G. 2014. "Constructing a Social Subject: Autism and Human Sociality in the 1980s." *History of the Human Sciences* 27 (4): 98–115.
- Jordan-Young, Rebecca M. 2011. *Brain Storm: The Flaws in the Science of Sex Differences*. Cambridge, MA: Harvard University Press.
- Kassraian-Fard P., C. Matthis, J. H. Balsters, M. H. Maathuis, and N. Wenderoth. 2016. "Promises, Pitfalls, and Basic Guidelines for Applying Machine Learning Classifiers to Psychiatric Imaging Data, with Autism As An Example." *Frontiers in Psychiatry* 7: 177.
- Keyes O. 2018. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–22.
- Keyes O. 2019. *Counting the Countless: Why Data Science is a Profound Threat for Queer People*. Cambridge: Real Life 2.
- Kitchin R. 2017. "Thinking Critically About and Researching Algorithms." *Information, Communication & Society* 20 (1): 14–29.
- Krafft P., M. Young, M. Katell, K. Huang, and G. Bugingo. 2019. Policy Versus Practice: Conceptions of Artificial Intelligence. Available at SSRN 3431304.
- Latour B. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. New York: Harvard University Press.
- Law J. 2004. *After Method: Mess in Social Science Research*. New York: Routledge.
- Lazer D., A. Pentland, L. Adamic, S. Aral, A. L. Barabási, D. Brewer, N. Christakis, et al. 2009. "Computational Social Science." *Science (New York, N.Y.)* 323 (5915): 721–723.
- L'heureux A., K. Grolinger, H. F. Elyamany, and M. A. Capretz. 2017. "Machine Learning with Big Data: Challenges and Approaches." *IEEE Access* 5: 7776–7797.
- Lugones M. 2016. "The coloniality of gender." In: *The Palgrave handbook of gender and development*, 13–33. New York: Springer.
- Maclure J. 2019. "The New AI Spring: A Deflationary View." *AI & Society* 35: 1–4.

- M'charek A., K. Schramm, and D. Skinner. 2014. "Topologies of Race: Doing Territory, Population and Identity in Europe." *Science, Technology, & Human Values* 39 (4): 468–487.
- McQuillan Dan. 2017. "Data Science as Machinic Neoplatonism." *Philosophy & Technology* 31: 253–272. <https://doi.org/10.1007/s13347-017-0273-3>
- Morandini J. S., A. Blaszczynski, M. W. Ross, D. S. Costa, and I. Dar-Nimrod. 2015. "Essentialist Beliefs, Sexual Identity Uncertainty, Internalized Homonegativity and Psychological Wellbeing in Gay Men." *Journal of Counseling Psychology* 62 (3): 413.
- Natale S., and A. Ballatore. 2017. "Imagining the Thinking Machine: Technological Myths and the Rise of Artificial Intelligence." *Convergence* 26: 3–18.
- Noble S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.
- Ortega F., and S. Choudhury. 2011. "'Wired Up Differently': Autism, Adolescence and the Politics of Neurological Identities." *Subjectivity* 4 (3): 323–345.
- Ortoleva P. 2009. "Modern Mythologies, the Media and the Social Presence of Technology." *Observatorio (OBS*)* 3 (1).
- Passi S., and P. Sengers. 2020. "Making Data Science Systems Work." *Big Data & Society* 7 (2). doi: doi:10.1177/2053951720939605.
- Roberts D. 2011. *Fatal Invention: How Science, Politics, and Big Business Re-create Race in the Twenty-first Century*. Princeton: New Press/ORIM.
- Rose N., and J. M. Abi-Rached. 2013. *Neuro: The New Brain Sciences and the Management of the Mind*. New York: Princeton University Press.
- Samuels E. 2014. *Fantasies of Identification: Disability, Gender, Race*. Vol. 10, NYU Press.
- Sato J. R., M. Q. Hoexter, P. P. de Magalhães Oliveira Jr, M. J. Brammer, D. Murphy, C. Ecker, and M. A. Consortium. 2013. "Inter-regional Cortical Thickness Correlations are Associated with Autistic Symptoms: A Machine-learning Approach." *Journal of Psychiatric Research* 47 (4): 453–459.
- Schultz S. 2019. Artificial Intelligence Detects a New Class of Mutations Behind Autism. *Princeton University* <https://www.princeton.edu/news/2019/05/28/artificial-intelligence-detects-new-class-mutations-behind-autism>.
- Seaver N. 2018. "What Should an Anthropology of Algorithms Do?" *Cultural Anthropology* 33 (3): 375–385.
- Seaver, 2019. "Knowing Algorithms." in *DigitalSTS: A Field Guide for Science & Technology Studies*, edited by Janet Vertesi and David Ribes. Princeton: Princeton University Press.
- Singh J. S. 2015. *Multiple Autisms: Spectrums of Advocacy and Genomic Science*. Minneapolis: University of Minnesota Press.
- Stark L. 2018. "Algorithmic Psychometrics and the Scalable Subject." *Social Studies of Science* 48 (2): 204–231.
- Terry J. 1999. *An American Obsession: Science, Medicine, and Homosexuality in Modern Society*. Chicago: University of Chicago Press.
- Thompson D. 2015. "What Lies Beneath: Equality and the Making of Racial Classifications." *Social Philosophy & Policy* 31 (2): 114.
- Tribune T. 2019. "AI Detects New Class of Genetic Mutations Behind Autism." *The Tribune* <https://www.tribuneindia.com/news/archive/health/story-779521>.
- Van der Ploeg I. 2012. "The Body as Data in the Age of Information." In: *Routledge Handbook of Surveillance Studies*, 176–183. Cham: Springer.
- Verhoeff B. 2013. "Autism in Flux: A History of the Concept From Leo Kanner to DSM-5." *History of Psychiatry* 24 (4): 442–458.
- Waidzunas T. 2015. *The Straight Line: How the Fringe Science of Ex-gay Therapy Reoriented Sexuality*. Minneapolis: University of Minnesota Press.

- Waidzunus T., and S. Epstein. 2015. "For Men Arousal is Orientation: Bodily Truthing, Technosexual Scripts, and the Materialization of Sexualities Through the Phallometric Test." *Social Studies of Science* 45 (2): 187–213.
- Wang Y., and M. Kosinski. 2018. "Deep Neural Networks are More Accurate Than Humans At Detecting Sexual Orientation From Facial Images." *Journal of Personality and Social Psychology* 114 (2): 246.
- Wolf C. T., and D. Paine. 2018. Sensemaking practices in the everyday work of AI/ML software engineering. *interface* .
- Woods R. 2017. "Pathological Demand Avoidance: My Thoughts on Looping Effects and Commodification of Autism." *Disability & Society* 32 (5): 753–758.
- Yang K. K., Z. Wu, C. N. Bedbrook, and F. H. Arnold. 2018. "Learned Protein Embeddings for Machine Learning." *Bioinformatics (Oxford, England)* 34 (15): 2642–2648.
- Zhou J., C. Y. Park, C. L. Theesfeld, A. K. Wong, Y. Yuan, C. Scheckel, J. J. Fak, et al. 2019. "Whole-genome Deep-learning Analysis Identifies Contribution of Noncoding Mutations to Autism Risk." *Nature Genetics* 51 (6): 973–980.
- Zimmer F., M. Stock, K. Scheibe, and W. G. Stock. 2019. "Fake News in Social Media: Bad Algorithms or Biased Users?" *Journal of Information Science Theory and Practice* 7 (2): 40–53.



Can artificial intelligence be decolonized?

Rachel Adams ^{a,b}

^aHuman Sciences Research Council, South Africa; ^bInformation Law and Policy Centre, Institute of Advanced Legal Studies, University of London

ABSTRACT

AI is altering not only local and global society, but what it means to be human, or, to be counted as such. In the midst of concerns about the ethics of AI, calls are emerging for AI to be decolonized. What does the decolonization of AI imply? This article explores this question, writing from the post-colony of South Africa where the imbrications of race, colonialism and technology have been experienced and debated in ways that hold global meaning and relevance for this discussion. Proceeding in two parts, this article explores the notion of de/coloniality and its emphasis on undoing legacies of colonialism and logics of race, before critiquing two major discontents of AI today: ethics as a colonial rationality and racializing dividing practices. This article develops a critical basis from which to articulate a question that sits exterior to current AI practice and its critical discourses: can AI be decolonized?

Introduction

In the totality of its machinic, computational, and imaginary manifestations, Artificial Intelligence is structurally, systematically, and psychologically altering not only local and global society, but what it means to be human, or, to be counted as such. Accounts such as James Williams' on the attention economy that detail how digital technologies function at a neurological level to capture and coax human impulses (2018); Brett Frischmann and Evan Selinger's analysis of how AI will not replace humanity but rather re-engineer us as computable (2018); and Shoshana Zuboff's work on surveillance capitalism that explicates a new world order where human behaviour has become the commodity of capitalist extraction (2018), all provide examples of how advanced technologies associated with AI are working at deep and complex levels in ways that are radically redefining what it means to be human.

Yet none of these texts, which constitute some of the leading work in the field, take into account the complex genealogy of intelligence: whose conception of intelligence is modelled within technology or how the idea has

been put to work in dividing people between the desired and the undesirable. Nor the history of the human body as machine and commodity borne from slavery and colonialism, such that Achille Mbembe names blackness as the prototype for the assemblage of the human–object of modernity (2017, 2019). Nor the ways in which the knowledges upon which AI is built – statistical enumeration of people and land – were advanced by imperial powers to control and contain colonial populations (Appadurai 1993; Breckenridge 2014; Kalpagam 2000, 2014; Said 1978), and led to the development of cybernetics and eugenics, as well as the idea that, through feedback mechanisms, both human and machine can be corrected and improved.

Indeed, for Mbembe we are encountering a third shift in the arrangement of race and blackness in global society: the first being slavery and colonization; the second being the development of writing and text, culminating in the formal processes of decolonization; and the third being the advent, proliferation, and ubiquity of digital technologies which represent the latest phase of high-modernity (2017). Similarly, for Aníbal Quijano, we are reaching a watershed moment in the global coloniality of power, with ‘the manipulation and control of technological resources of communication and of transportation in order to impose the technocratization/instrumentalization of Coloniality/modernity’ (2017, 364), together with ‘the mercantilization of subjectivity and life experiences of individuals’ (2017, 365).

AI, too, is in the midst of reconciling three forms of discontent,¹ whose relation to the forms of coloniality and the historical construction of race at work in the world today has yet to be fully understood. First, manifest racial and gender bias within AI technologies (Benjamin 2019; Buolamwini and Gebru 2018; Keyes 2018; Noble 2018). In June 2020, E. Tendayi Achiume, United Nations Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia, and related intolerance, issued a report to the United Nations Human Rights Council that found ‘emerging digital technologies exacerbate and compound existing inequalities, many of which exist along racial, ethnic and national origin grounds’ (paragraph 4). The second discontent, echoing Quijano’s description of the mercantilization of life, regards the way AI and digital technologies are globally commoditizing human experience (Zuboff 2018). Within this data paradigm, the human is substituted as an assemblage of their data points which are, in turn, taken as a sign of the real (Baudrillard 1994). The third discontent is geopolitical, pertaining to the emergence of an arms race characterized by the ambitions of transatlantic nation states to be ‘global leaders’ (Vladimir Putin, cited in Vincent 2017) in AI innovation (Garvey 2019).² This, together with the second discontent, is central to the ideas of ‘data colonialism’ and ‘digital colonialism’ naming, within current AI discourse, the brazen flurry to extract and exploit personal data and data systems (Birhane 2019; Couldry and Mejias 2019; Crawford et al. 2019; Ricaurte 2019; Mhlambi 2020). It relates as well to Ian Hogarth’s critique of ‘AI

nationalism' where new dependencies are being tacitly enforced between low-tech and advanced-tech states and are set to follow the historical divide of Global North/South (2018). Perhaps more critically, the drive to dominate in the production and use of AI is revealing of the project's hegemonic impulses, and the neo-Darwinian linearity of the evolution of science which will 'leave behind' those who do not conform to catching up. In light of these concerns, Mbembe's provocation to critique the apparatuses of race and blackness in digital technologies, and Quijano's warning against the hegemony of technocratization which encompasses, at once, the consolidation of modes of commodification of the human-object, becomes ever more urgent.

The foremost effort to resolve (or reconcile) the discontents within AI have taken shape within discussions around, and the evolution of, ethical principles to govern the development and application of such technologies (Ulnicane et al., this issue). The overall agenda, it appears, is to create benevolent AI. The idea of decolonization has emerged largely *within* the discourse of AI ethics, in recognition of the racialized and imperialistic effects of the field's power (Bell 2018; Mohamed 2018; Birhane 2019; Peña and Varon 2019; Mohamed, Png, and Isaac 2020). Examples include, 'Decolonizing AI' listed as a theme under the broader project on AI Narratives and Justice at the University of Cambridge's Leverhulme Centre for the Future of Intelligence,³ and Genevieve Bell's (2018) lecture titled 'Decolonizing AI,' where she mobilized the term as a critical tool for reading power and control back into the history of AI and complexifying the sociology of the discipline. And, more recently, Shakir Mohamed, Marie-Therese Png and William Isaac have published their vision of 'decolonial AI,' advocating, in particular, for decoloniality as a tactic for engaging in what they name 'sociotechnical foresight' to 'support future technologies that enable greater well-being, with the goal of beneficence and justice for all' (2020, 1).⁴

To date, the idea of decolonization has been mobilized within the field of AI in a way that asks: What can decolonization mean *for* AI? That is, how can the critical ideas of decolonial thought be applied to and utilized to broaden and critique the field of AI? In this article, I explore what I argue to be a necessary shift in perspective, and ask instead: What does AI mean *because of* colonialism? It is necessary because, without thinking *within* the decolonial paradigm which demands a critical positioning of AI amidst the historic totality of colonialism and race, we (as critical thinkers of AI) run the risk of reproducing the very problematics decoloniality seeks to disrupt. Thus, the first section of this article provides a brief exegesis on what the notion of decoloniality has come to mean with a focus on its ontology within Africa, and explores how decolonization in the context of AI research encompasses an imperative to reveal, critique, and radically unbalance both the legacies of colonialism *and* the logics of race instituted by colonialism that are at work within the field. While I engage with ideas of decoloniality and histories of colonialism in various parts of the

world, the emphasis within this article is on the African region, where the inter-connections between the legacies of colonialism and the logics of race have been well evidenced and received much attention within critical thought. In the second section, I advance a genealogical analysis of AI's participation within two themes: first, ethics as colonial rationality, and second, what I call the dividing practices of coloniality and race which produce a 'world of apartness' (Madlingozi 2018). These critical histories allow for an understanding of both how and why the emergence of practices of coloniality and pathologies of race within AI today are taking place, and provide the critical basis from which to imagine an alternate future.

The contribution made here is, then, twofold. First, to begin to articulate why it is not sufficient to simply identify the reproduction of racializing and colonialist logic in the science and practice of AI today; rather, what decolonial thought demands is to show – precisely – how and why AI as a field *depends on*, and was made possible by, these logics. And second, to pose the question of whether a decolonial AI future that makes space for multiple and culturally varied accounts of intelligence and being is indeed possible. In short, can AI be decolonized?

Decolonization/decoloniality: a brief exegesis

Decolonization is a contested term. In its formal historical sense, decolonization refers to the process of transferring legal, administrative, and territorial power from colonial hands to indigenous local governments, and thus the establishment of modern nation states independent from European empires (Jansen and Osterhammel 2017). While the abrogation of colonization as a mode of forced rule took place in countries across Africa, Asia, the Americas, and Oceania over the course of the twentieth century, the legacies of colonialism and the racial bifurcation that it was both built on and produced, were not so readily dismantled and overcome. From the ashes of formal colonial rule, ulterior forms of imperialism arose, undermining the sovereignty of post-colonial states and peoples through economic and legal dependency, epistemic authority, exploitation and social abjection. The enterprise of decoloniality (what we can call, distinct from decolonization, 'decolonialization' (see further Mabhena 2017)) seeks, precisely, to identify, critique and undo the ulterior forms of imperialism within contemporary global society that were instituted and made possible by the colonial project.

The notion of decoloniality has been refracted in various local contexts. In South America, ideas around decoloniality arose in the 1990s, led by the work of Aníbal Quijano, Ramon Grosfugal and Walter D. Mignolo, amongst others. Here, coloniality names the Janus-face of modernity and capitalism: colonialism is the central force that makes possible the interlocked projects of modernity and capitalism. Thus, the end of modernity is 'the ultimate

decolonial horizon' (Mignolo and Walsh 2018, 4). For Quijano, *decoloniality* constitutes the inverse force of coloniality and not, therefore, a form of deconstructive power that comes *after* colonialism. In this sense, Quijano appears to draw from Foucaultian thought on power/resistance to conceive of 'de/coloniality' as the potentiality to undo and invalidate, that has always been, and will always be, caught up in the active presence of what he names 'the coloniality of power' (2010). Put differently, 'de/coloniality' names the coexistence of the 'coloniality of power' *with* the possibility of its unravelling. Identifying the signifying practices (the ways in which authoritative meaning is (re)produced) of the coloniality of power – the epistemic, cultural, political and economic apparatuses through which the oppression of (sub)alterity is constituted and maintained, and Eurocentric ways of knowing, being, and thought are reproduced as superior and singularly legitimate – is globally critical to the decolonial project.

Within decolonial thought from the African region, race is the chief organizing principle (Ndlovu-Gatsheni 2015), the central figure within the entanglement of coloniality and being. Indeed, in South Africa,⁵ where scholars have sought to engage with decoloniality in delineating the vestiges of colonialism that continue to shape experiences of subjugation and oppression along racial lines today, Tshepo Madlingozi has articulated that 'the entrenchment of a world of apartness' constitutes one of its major legacies (2018).⁶ Formalized in apartheid South Africa, this 'world of apartness' continues globally through latent and overt systems and structures of racial (and gendered) bifurcation that mark, divide, categorize, classify, and hierarchize individuals according to social norms of un/desirability. More broadly, the 'world of apartness' is the product of the historic totality of race and racism. Crain Soudien (n.d.) narrates the global history of race along three major arrangements, beginning with its 'invention' as a myth of European superiority, which peaked in the seventeenth Century with the transatlantic slave trade. During the second historical arrangement over the nineteenth and early twentieth century, race entered scientific discourse and was legitimated as empirical knowledge with race science and, ultimately, projects such as eugenics. Critical to race science was the development of scientific discourses which sought to empiricize differential racial attributes, whether physiological, such as phrenology, or cognitive, such as intelligence tests.⁷ In the third arrangement, from the 1930s onward, race became recognized as a social and ideological construct, and race science was slowly discredited (Soudien n.d.). Decoloniality, then, could be described as the final arrangement of race, as a radical anti-racism to come. Critically, this typology points to the centralizing performance of Western reason (encompassing thought, knowledge, myth, or other shared codes through which a particular value-laden logic is established and operates) in the production and continuation of the idea of race and its attendant pathology: racism. Science was deployed to justify myth and, ultimately, conquest.

Following Frantz Fanon, African thinkers such as Ngũgĩ Wa Thiong'o (1986) and Chinweizu (1987) came to critique the governing power of Western

epistemologies over colonized populations, and called for the ‘decolonization of the [African] mind’ as an essential condition for emancipation. Wa Thiong’o discerned how language works as a hegemonic vessel of knowledge and signification through which Eurocentrism continues to hierarchize the world and devalue all things African. He writes: ‘language carries culture, and culture carries, particularly through orature and literature, the entire body of values by which we come to perceive ourselves and our place in the world’ (1986, 16). Decolonizing the mind, therefore, necessitated for wa Thiong’o a re-embrace of African languages in particular,⁸ and African value systems more generally, in order to contest and provide alternatives to Eurocentric ways of knowing and being in the world.⁹ Critically, the African canon on decoloniality calls not for the negation of Western modes of thought or being, but for the radical pluralization of the world and what it means to live together as humanity within it.¹⁰

Within the United States (and elsewhere (Adebisi 2019)), decolonization has, at times, been adopted as a stratagem for naming and addressing broader issues of social justice concerning systemic and structural forms of racism, discrimination, and oppression. This has been critiqued by Tuck and Yang as abstracting decoloniality for easy adoption as a metaphor (2012). They argue that this both detracts from the actual project of decolonization (which in the United States pertains critically to the ‘repatriation of Indigenous land and life’¹¹ (2012, 1) and re-performs the particular settler logic of appropriation. Indeed, to produce decoloniality as a metaphor is both to seize the epistemic strategic of resistance (or as Walsh and Mignolo name it ‘re-existence’ (2018, 3)¹²) and to reduce it to rhetoric, to a figure of speech which, in the chain of signification, is further removed from the actuality which it seeks to describe and deconstruct. In short, it subverts decolonization and neutralizes its potentiality. They conclude: ‘when metaphor invades decolonization, it kills the very possibility of decolonization; it recenters whiteness, it resettles theory, it extends innocence to the settler, it entertains a settler future’ (2012, 3).

The key difference between the articulations of decoloniality in South America and the African region as described, and the critique set out by Tuck and Yang, is that decoloniality *is* the strategy, a situated and responsive strategy to the particular form of *power over* that coloniality impressed upon the world, and not an autonomous and separable concept that can be put to work as a strategy *for* something else. Decoloniality must be revolutionary, South African scholar Sabelo Ndlovu-Gatsheni avers, as anything less is reformism and ‘another way to seem progressive’ (2020). As a revolutionary strategy, decoloniality aims toward the radical transformation of the Western symbolic order of the world. For Joel Modiri, while decoloniality is historically situated, this does not confine it to a finite prescription or task. Echoing the insurgency of Ndlovu-Gatsheni, Modiri writes that ‘decolonization is an insatiable reparatory demand, an insurrectionary utterance, that always exceeds the temporality

and scene of its enunciation. It entails nothing less than an endless fracturing of the world colonialism created' (2019). Indeed, Soudien's typology of race makes this clear: the myth of race and its attendant racism were central conditions which made the emergence of colonialism possible. Thus, race goes beyond colonialism; therefore, the decolonial project transcends the deconstruction of coloniality.

What does it mean, then, to talk about decolonizing AI? I take this to comprise of two questions. First, What does the idea of decolonization, and the invocation to decolonize, entail? To restate my answer, given above, to decolonize now constitutes a distinct injunction from formal decolonization. While the latter is bound by a finite temporality of state sovereignty, the former requires a radically more distributed effort across multiple planes of meaning to fracture both histories and futures which delimit the participation of colonized peoples and life forms in the global humanity of life. Second, to which I now turn: What does the injunction to decolonize AI require us to do? This question is complex and demands a continually reflexive critique of the conditions of possibility for life and living – now, then, and in the future – being fixed and bounded by AI and its attendant discourses. However, in view of the above, this second question requires attention to the following: to identify, and strive to undo, the legacies of colonialism, including the rationalities that allowed it to be accepted, *and* the logics of race at work in the idea and practice of AI today. In what follows, I attend to these issues through an examination of two major themes and their historical relation to AI: ethics as a colonial rationality, and the production of the 'world of apartness' (Madlingozi 2018) through the dividing practices upon which the West claims and pursues its superiority – including, in particular, the concept of intelligence.

Ethics and the rationality of Empire

Computer scientist has Timnit Gebru stated that ethics is the 'language du jour' in AI discourse (2020; see also Ulnicane et al., this issue). Discursively, it is posited that through advancing ideas such as 'AI for Good,' 'Fair and Responsible AI' and 'AI for Humanity' (in France and Canada), particularly by incorporating these aspirations and values within normative frameworks for ethical AI, discontent within the field can be addressed and resolved. This has met with some criticism. Pratyusha Kalluri, for example, has pointed out that "fair" and 'good' are infinitely spacious words that any AI system can be squeezed into,' and aptly stresses that the question AI ethics should be examining is how power works through such systems and to what effect (2020; see also Crawford et al. 2019). In addition, Greene, Hoffmann, and Stark (2019) have emphasized how AI ethics assumes a universality of concerns which can be objectively measured and addressed, summarizing the assumptions upon which the discourse is based as follows:

- (a) the positive and negative impacts of AI are a matter of universal concern,
- (b) there is a shared language of ethical concern across the species, and
- (c) those concerns can be addressed by objectively measuring those impacts (2126).

‘This,’ they write with irony, ‘is a universalist project that brooks little relativist interpretation’ (2019, 2126).

This is where the language of decoloniality has proffered some new thinking. In their piece on ‘Decolonial AI’, Mohamed, Png, and Isaac (2020) have, amongst other propositions, advocated for dialogue between the AI metropolises and peripheries as a means of developing ‘intercultural ethics’ (17). Specifically, they write that dialogue can facilitate ‘reverse pedagogies’ wherein the metropolises can learn from the peripheries, and that ‘intercultural ethics emphasizes the limitations and coloniality of universal ethics – dominant rather than inclusive ethical frameworks – and finds an alternative in pluralism, pluriversal ethics and local designs’ (2020, 17). Sabelo Mhlambi has taken this a step further by developing a framework for AI ethics based on the Nguni philosophy of Ubuntuism (2020). Critiquing Western reason of rationality in shaping the philosophical terms on which AI and AI ethics is dominantly conceived, Mhlambi details how the Sub-Sahara African notion of Ubuntu, which centres on the relationality of personhood, can undergird a framework for addressing the two major challenges in AI: surveillance capitalism and data colonialism (2020).

While this is important, the emerging discourse around ethics and decolonizing AI has yet to develop critical thought around the idea of ethics itself. As above, Mohamed, Png, and Isaac briefly note ‘the limitations and coloniality of universal ethics’ (2020, 17), but it is critical to understand precisely why the dominance of this particular version of ethics – vested as it is in the history of Eurocentric thought around morality, legality/governance and personhood – is so problematic, and what the effects might be of uncritically drawing decoloniality *into* this discourse. Put differently, should decoloniality be subsumed as a new tool for AI ethics, without critique of the way in which the idea of ethics has been historically put to work in rationalizing colonial practices (see Mbembe 2017, 12; Spivak 1988, 9), it runs the risk of not only appropriating decoloniality as an abstract metaphor, as Tuck and Yang (2012) warned against, but also of reproducing the very logics of race that colonialism instituted. Let us now more closely examine this problematic formulation, ‘AI ethics.’

In 2019, a study was published in *Nature* identifying over 84 ethical standards for the use and development of AI developed globally in the last five years (Jobin, Ienca, and Vayena 2019; Ulnicane et al., this issue). Despite being titled ‘The Global Landscape of AI Ethics Guidelines,’ amongst these 84 AI ethics standards, none listed are from the African continent or even the Global South. Most were developed in the United States, UK or by

international institutions. Mohamed, Png, and Isaac (2020) similarly note how national AI policies or strategies are almost exclusively found in the Global North, and where efforts to develop a national policy around AI are arising in countries within the Global South, this is being driven by supra-state bodies such as the World Economic Forum. As ethical benchmarks, these standards are paternalistically positioned as universal: applicable for all, everywhere. In addition, the scientific practice of promoting ethical AI through strengthening or testing the ‘fairness’ of AI systems (the extent to which they exhibit social biases, in particular) performs a similar conceit in presuming the scene of the Global South – or more specifically in this case, the African region – to be a place where ‘ethics,’ as such, is yet to be fully established. Now a well-documented case (Ballim and Breckenridge 2018; Arun 2020; Arun 2020), in 2018, when the issue of racial bias and the non-recognition of Black faces by AI-driven facial recognition technologies was peaking following the work of Buolamwini and Gebru (2018),¹³ a Chinese facial recognition company signed a deal with the Zimbabwean government for access the records of the national population registry, which contained facial imagery of millions of Zimbabweans, to train the company’s algorithmic technologies to better recognize Black faces. By reducing the potential for bias, the system would ultimately be more ethical. While Ballim and Breckenridge (2018) condemn this incident for exploiting the inadequate data protection provisions in Zimbabwean law, it is not all that different from the practice of beta-testing newly developed AI systems in African countries (Mohamed, Png, and Isaac 2020). Calling it ‘ethics-dumping,’ Mohamed, Png, and Isaac point to the notorious company Cambridge Analytica as exemplar, in that it developed algorithmic systems for use in the US and UK by beta-testing them in Nigeria and Kenya (2020, 11). This follows the now centuries-old colonial conceit of what Jan Smuts euphemistically called the ‘laboratory of Africa’ (1930), where the collateral damage of scientific advancement could be safely externalized to places and people considered expendable (see Bonneuil 2000; Tilley 2011; Taylor 2019). Moreover, the epistemological foundations of AI cannot be extricated from Francis Galton’s work in the development of statistics – particularly on inference, regression, correlation, and the normal distribution curve – which arose out of his explorations in Southern Africa, where he applied his statistical science to native populations in order to measure human differences and intelligence (Breckenridge 2014, Chapter 1).

In these instances, the idea of ‘ethics’ is situated as the supreme value of the Occident, to be proselytized on the Africa region, which is, in turn, and in relation to the ‘ethical West,’ positioned as ‘pre-ethical’ (Mbembe 2017, 49) – as a world apart. Indeed, that Europe believed itself to be ‘helping’ and ‘protecting’ its African colonies constituted the central creed of the civilizing mission of colonialism (Césaire 2001); as Spivak reminds us, ethics ‘served and serves as [its] energetic and successful defense’ (1988, 5). Yet as ethics was put to work

to justify both the civilizing mission of colonialism and the utilization of Africa as a laboratory for Western scientific progress, it enacted another conceit of the colonial order of things: that Western reason is neutral, universal, and objective; that it could be dislocated from the context in which it arose and applied elsewhere. Positioned as a ‘point zero’ (Santiago Castro-Gomez unpublished work, cited in Grosfugal 2011, 6) from which to survey the world, Western knowledge and rationality claimed ascendancy as the only real way of knowing and understanding the world. This is a critical problematic within decolonial thought (Grosfoguel 2007; Ndlovu-Gatsheni 2013), and a central assumption within AI: that intelligence and the production of knowledge can be outsourced to a machine presupposes such knowledge to be both separable from the context in which it was produced and applicable to other contexts and realities.

Dividing practices

The production of ‘the world of apartness’ (Madlingozi 2018) takes place through what I am calling ‘dividing practices.’ In this section I explore briefly the provenance of systems of enumeration, quantification, and classification within colonialism, and the ways in which AI reproduces the divisive logics of race, before turning to critique the notion of intelligence in particular. I take the term from both Michel Foucault who, writing on the production of the objectivization of the subject, speaks of ‘dividing practices’ which divide the subject from others and within itself (1982, 777–778), as well as Edward Said’s critique, set out in *Orientalism*, of the dividing line – discursively formed – between the Occidental and Oriental worlds, which the former ‘paradoxically presupposes and depends on’ (1978, 336). For both, power resides with those who can make the catechistic decision to divide.

It is well-noted how AI systems sort personal data according to socially ascribed normative markers. At times, these markers are directly racialized or gendered (Keyes 2018; see also Keyes, Hitzig, and Blell, this issue), such as a system that only allows women access to a female changing room (Ni Loideain and Adams 2019). Other times, these markers may be implicitly biased, such as systems for targeting advertising and policing based on postal codes (Benjamin 2019). These systems classify, sort, and rank personal data through processes of data collection, curation, and annotation, using advanced statistical methods for modelling distribution and measuring correlation (as first developed by Galton) in order to calculate risk, predict behaviour, and optimize the systems’ own functions. In these contexts, data assemblages constitute a representation of the individual that are taken (by commercial and state power) as a sign of the real (Baudrillard 1994). Moreover, as Birhane (2019) has pointed out, these systems of abstract representation work to further marginalize those who do not fit the ‘data-type’. Indeed, Quijano (2017) has spoken of

modern systems that function through identifying and classifying individuals as fundamentally ‘de-equalizing’, presumably as the application of these practices to human subjects supposes a fixed, *a priori* and quantifiable difference, the social-construction of which is forgotten.

As noted above, much has been published about how these systems reproduce social biases (see also Holzmeyer, this issue) with many accounts noting the racial logic and imperial power at work (Buolamwini and Gebru 2018; Keyes 2018; Noble 2018; Benjamin 2019). However, rather less examined in relation to practices of AI today, is the way in which these statistical systems were developed and appropriated within former colonies to control and divide colonial subjects.¹⁴ Indeed, Said wrote that, ‘rhetorically speaking, Orientalism is absolutely anatomical and enumerative: to use its vocabulary is to engage in the particularizing and dividing of things Oriental into manageable parts’ (1978, 72). Similarly, in narrating the enumerative practices of colonialism in India – which he critiques as having both a disciplinary and pedagogical effect, in delimiting colonial subjectivity and in training colonial administrators respectively – Appadurai writes:

The link between colonialism and orientalism [...] is most strongly reinforced [...] at the loci of enumeration, where bodies are counted, homogenized, and bounded by their extent. Thus the unruly body of the colonial subject (fasting, feasting, hook-swinging, abluting, burning, and bleeding) is recuperated through the language of numbers that allows these very bodies to be brought back, now counted and accounted, for the humdrum projects of taxation, sanitation, education, warfare, and loyalty. (1993, 334)

Enumeration and the production of statistical knowledge in the colonies performed a number of functions, including entrenching and policing colonialist binaries of colonizer/colonized and their derivatives, but also in enforcing divisions between colonial populations,¹⁵ and as a form of remote colonial rule. On both a structural and individual level these colonial archives functioned as a kind of palimpsestic¹⁶ form of abstract representation that were taken as a token of ‘radical realism’ (Said 1978, 72): a fixing of the ontology of the colonies and its people by Western knowledges, just as the data assemblages of today work to fix individuals by taking their data as a sign of the real. Writing of forms of representation at work within systems of racism, Mbembe speaks of a ‘will to representation [which] is at bottom a will to destruction aiming to turn something violently into nothing’ (2019, 139). In this way, to constitute something in the form of something else – *something* more manageable and more malleable to forms of racializing power – consists of an essential and violent erasure of the original. Imperial knowledge practices based on abstract and racialized representations constituted not only a way of dividing the self from others and from itself, but worked to erase those who fell on the wrong side of the dividing line through substituting them with their representation.

Comparably, Simon Gikanda chronicles the slave masters' fastidious recording-keeping of the actions of their slaves, such that this archive constituted the evidence of the latter's objectification: 'as chattel, as property, and indeed as the symbol of the barbarism that enabled white civilization and its modernist cravings' (2015, 92). That Simone Browne now writes of data-driven surveillance systems being put to work to surveil and bind Black lives in particular, as exacting the self – its body and behaviour – to testify as evidence against itself, holds then, a critical provenance within the history of the colonial management of blackness. The effects of these systems, such as AI-enabled biometric technologies in public spaces, which Browne describes as reifying structures for racial difference, is to produce an 'ontological insecurity' – an alienation within, or a dividing practice of, the racialized self (2015, 109).

Some of the most advanced biometric systems in the world today utilize facial recognition technologies within their technological make-up. These technologies function by reading the signatures of human faces, such as the distance between facial features, and comparing the image to a database of facial imagery in order to detect physiognomic patterns that correlate to the social status – gender, race, age, sexuality (Keyes 2018; Keyes, Hitzig, and Blell, this issue) – of the person being 'recognized.' As these practices enforce social stratification and re-inscribe racialized hierarchies, they also repeat the very logics of race and racism: to deduce from signs and surface appearance who an individual is and what they can do and be in this world. Race science legitimated this logic through the production of knowledges that sought to demonstrate the link between surface appearances – skin colour, facial features, skull size – and inherent cognitive abilities and behavioural traits (Soudien n.d.). Today's facial recognition technologies employ similar practices in measuring facial diameters and expressions as a means of inferring intent, predicting behaviour (Chinoy 2019), and even understanding intelligence (Qin et al. 2016). More critically, these systems provide the tools that allow for the return of race science under the guise of securitization, market efficiency, and risk management.

From intelligence to reason

'It seems to us that in intelligence there is a fundamental faculty, the alteration or the lack of which, is of the utmost importance for practical life. This faculty is ... the faculty of adapting one's self to circumstances' (Alfred Binet, cited in Chollet 2019). Thus reads the epigraph to the section in AI researcher Francois Chollet's recent paper that sets out how advances in AI systems can be assessed through a psychometric test apt for measuring his revised, actionable definition of [human-like] intelligence. Chollet's central claim is that in order to advance development toward artificial general intelligence,¹⁷ the field must delineate a precise and measurable notion of human intelligence against which to

benchmark progress. He posits the utilization of a psychometric test that can measure the abstraction and reasoning capabilities of AI systems, which he insists are the key attributes of intelligence. The results of these tests can then be appropriate to best determine the optimal curricula from which the system can be improved. Chollet asserts that:

We posit that the existence of a human-level ARC [abstraction and reasoning corpus] solver would represent the ability to program an AI from demonstrations alone (only requiring a handful of demonstrations to specify a complex task) to do a wide range of human-relatable tasks of a kind that would normally require human-level, human-like fluid intelligence. As supporting evidence, we note that human performance on psychometric intelligence tests (which are similar to ARC) is predictive of success across all human cognitive tasks. (2019, 53)

The optimum AI system, then, displays the highest form of human-like intelligence in reasoning and abstraction and is disposed to self-correct with minimal intervention and exposure to curricula. For, as he asserts, [human-like] ‘intelligence lies in broad or general-purpose abilities; it is marked by flexibility and adaptability (i.e. skill-acquisition and generalization), rather than skill itself’ (2019, 27), and ‘a fundamental notion in psychometrics is that intelligence tests evaluate broad cognitive abilities as opposed to task-specific skills’ (2019, 13).

Offering a history of the concept of intelligence in European thought through Aristotle, Hobbes, Locke and Rousseau, Chollet’s well-cited paper stands as a disquieting testimony to AI’s celebration of decisively Western notions of intelligence (see also Collins; Blackwell, this issue). Like ethics, intelligence is a value-laden concept which has historically operated as a racialized dividing practice to differentiate between peoples and reaffirm white superiority (Cave 2020). Indeed, Alfred Binet, of Chollet’s epigraph, was the pioneer of psychometric and intelligence tests who originally established, as Cave notes, ‘the notion of “mental” age in order to determine which children were so behind that they should be given special education’ (2020, 31). Such tests went on to be used in the US and British colonies to justify the idea that intelligence was a hereditary attribute largely endowed to the white race (Sehlapelo and Blanche 1996; Tilley 2011; Laher and Cockcroft 2014). As reported in an American education bulletin, a 1916 study of South African Zulu children described finding that aspects of what was by then the Binet-Simon test ‘requiring memory and observation were readily answered [by the Zulu children], but that those requiring abstract thought were seldom answered’ (Martin 1916, 143). Within the context of these tests and the Western history of intelligence, abstract reasoning was considered a higher form of cognitive ability than memorization. In this case, the reporter took caution to note that the test had been modified so as to more fairly take into account cultural differences – a trend which was to continue throughout the evolution of psychometric tests,

particularly in culturally diverse places like South Africa, despite still fundamentally serving to highlight differential cognitive proficiencies in races, regardless of whether this was analysed as hereditary or environmental (Sehlapelo and Blanche 1996; Laher and Cockcroft 2014).

The legacy of the original Binet test in the various evolutions of psychometric and intelligence tests that followed emphasized how such tests could be used to determine not only cognitive-appropriate educational interventions, but the essential educability of individuals. By 1950s and 60s South Africa, instruments testing educability became used to test the adaptability (as foreshadowed by Binet in the quotation above) of ‘the natives’ to new working demands (Laher and Cockcroft 2014). Following the publication in 1943 of his book, *African Intelligence*, Simon Biesheuvel led a proposal to develop a psychological assessment to, amongst other objectives, determine ‘the extent to which (the African’s behaviour) is modifiable’ (1958, 162). As Laher and Cockcroft note, ‘this group of psychologists felt it necessary to “understand” the African personality in order to justify the inferiority of Africans to other races, and to control and modify African behavior’ (2014, 307). Despite taking place well after eugenics had been dismissed from the global scene as an unethical science and practice, the hallmarks of the idea that individuals, exposed to the right stimulus, could be corrected and improved, remained. Indeed, as Cave has pointed out, intelligence tests were central to the eugenicist ideology and project (2020, 31). Stepping remarkably close to the logics of eugenics, Chollet asserts the emergence of human cognition as evolutionary (2019, 22) and, on this basis, advocates for AI systems to be developed in environments that are optimally designed to promote its effective advancement. Indeed, one of the primary assumptions of AI, which Chollet’s paper aims to improve upon, is the idea that human intelligence can be anatomized and known in such detail that it can be replicated and simulated by machines. That is, it presumes the total knowability of the human. And, in this way, AI aligns with the fundamental premise of cybernetic and eugenic reason, appropriated too by Biesheuvel in 1950s South Africa, that with precise knowledge coupled with exacting interventions, the human can be corrected, improved, and even – as AI advocates claim – replicated.

Cave upholds that AI’s uncritical promotion of Western notions of intelligence may underlie the lack of diversity within the field, as well as the fear that ‘superintelligence,’ once attained, will colonize the human species writ large (2020; see also Falk, this issue). However, perhaps more critically, the ways in which the notion of intelligence has been put to work historically as a centralizing principle within the field of eugenics and has a history, now re-emergent, as a dividing practice with immediate racializing effects, gives credence to Mbembe’s critique of blackness as the prototype to the modern human-object, and demands a rethinking of the politics of intelligence today.

Conclusion: after worlds

Decolonial thought is far more than a tool to problematize AI. It is an invocation to make intelligible, to critique, and to seek to undo the logics and politics of race and coloniality that continue to operate in technologies and imaginaries associated with AI in ways that exclude, delimit, and degrade other ways of knowing, living, and being that do not align with the hegemony of Western reason. It is located and specific. It is about the production of race and divided worlds; it is about power and the precise effects of power on being in the world today; it is about knowledge and how knowledge is ascribed legitimacy and value; and it is about a politics of resistance that enters and undoes the object of its critique. This includes, as I have outlined above in relation to ethics in particular, the discourses that rationalize and obscure the history and effects of AI. In addition to those explored here, there are many other ways in which decoloniality must be brought to bear on the field and practice of AI,¹⁸ such as the invisibilizing labour that sustains the industry, its utilization of traditional gender binaries within systems and products (Adams 2019), the biopolitical intersection of gender and race in the anthropocentric production of machines, and the links between AI and contemporary modalities of capitalist modernity. Indeed, much more work is needed to fully understand the entanglement of AI with coloniality and the pathologies of race.

However, in drawing on the discourse of decoloniality, critics of AI must resist the sublimation of decoloniality as another rationality that justifies and legitimates AI. To do so re-performs the very abstractions and disembodiment of thought that decoloniality seeks to resist. To be clear, if the decree to decolonize AI is not addressing race and historically embedded forms of Occidental *power over*, nor seeking to rupture the epistemological and teleological assumptions of the discipline and related fields from within a historical reading of their formation and appropriation within colonial regimes, then decolonization is being misappropriated as a metaphor, and its usage in the discourse has become a part of that to which decoloniality proper must address in its critique. This becomes ever more critical where surveillant AI technologies are being used to thwart decolonial resistance to racism and neo-imperial power (Ndlovu-Gatsheni 2020).

Further, it does not go far enough to restate that AI is having a racializing effect, or that its ubiquitous power throughout the world is hegemonic and neo-imperialistic. If colonial modes of power over and dividing practices of racism are being re-instituted through AI behind the veil of technocracy, what is the precise form of this re-institution of race and colonialism? How can AI be located within the *longue durée* of colonialism and race? Through examinations of the critical histories of the assumptions upon which the field is based and the knowledge practices in which it engages, a new understanding

of its effects of, on, and through power can emerge which can create the space for other localized and culturally diverse ways of understanding and doing AI, such as those being explored by Alan Blackwell (this issue). However, these critical histories, as briefly set out above, also point to another consternation: Rather than reproducing the logics of race and reaffirming the legacies of colonialism, AI depends upon them. Recall Said's prescription that the dividing line between the Occidental and Oriental worlds was 'paradoxically presuppose[d] and depend[ed] on' by the West (1978, 336). As such, the task of decolonizing AI requires, as I have argued here, a critique of the ways in which AI is made possible by, and depends upon, colonial forms of power and the dividing practices of racialization. But Said intimates another site of meaning by suggesting that the presupposition and dependence of Western power on racialization and colonialism is paradoxical, thus leading to an impasse to which Western reason has no answer. Within AI, these aporias can be glimpsed in industry's dogmatic pursuit of simulated intelligence, which latently affirms that intelligence is not hereditary but environmental; and in the industry's manipulation of behaviour, attention, and thought, indicating the breakdown of the autonomous Cartesian self which, paradoxically, ideas about intelligence and automation in AI are modelled on. For postcolonial thinkers, it is precisely in these impassive moments that the Western equation of thought becomes unbalanced, that a politics of speaking from the South can regenerate as productive and life-affirming (Spivak 1988; Mbembe 2017), thus articulating the otherwises and elsewheres of 'a humanity made to the measure of the world' (Césaire 2001, 73).

Indeed, recognizing that the undoing of coloniality was always a possibility of the original event of conquest – that it always could have been otherwise – simultaneously reinforces the possibility of different futures to come. With the global acceleration of AI and its supporting discourses, it is becoming increasingly difficult to imagine a future in which it is not dominant. Within these imagined future scenarios, AI constitutes another step in the evolution of humanity's triumph over the world. Contained within this imaginary are neo-Darwinian notions that those left behind by the technological revolution are not worthy of the new world. If it risks leaving so many behind, can AI, as currently imagined, ever be 'good,' 'benevolent,' or 'decolonial'? To speak of decolonizing AI not only then contains the imperative to collectively reimagine a multifarious world space and ask whether AI can be ascribed a role within, and conducive to, this new imagining – but also to be imaginative enough to conceive of a future without AI.

Notes

1. The concerns outlined here do not constitute the totality of critiques being levelled against AI, but pertain, particularly, to those with most direct relevance to the

- arguments under discussion here. Other concerns include those around transparency and accountability (for a critical account see Adams 2020), and labour (see Weinberg 2019; Crawford and Joler 2018), as well as other discontents outlined and critiqued in this issue.
2. Within policy documents relating to AI or the Fourth Industrial Revolution, the UK, US, Denmark, Germany, Finland, Italy, Japan and South Africa all note an ambition to become (or maintain the position of, in the case of the US and China) the (or 'a', for South Africa) global leader in AI. Yet this is history rhyming if not repeating itself: Garvey (2019) shows how this global competition is not, in fact, a new phenomenon, but the second 'AI arms race.'
 3. <http://lcfi.ac.uk/projects/ai-narratives-and-justice/decolonising-ai/>.
 4. I am looking here broadly at the idea of decolonisation as it is raised in the context following formal decolonisation of largely African states. I recognize that there is much work taking place in indigenous communities in building indigenous forms of AI and machine and deep learning, which draw on the terminology of 'decolonization' within their objectives. See, for example, <https://www.indigenous-ai.net>. However, the ways in which AI is being used as *a tool for* decolonisation is a question beyond the focus of this piece.
 5. The notion of decoloniality has surfaced in recent discourse and social movements in South Africa around the public university, and the premise that while formal decolonisation had taken place, universities and knowledge production continued to be fashioned within colonialist epistemologies which favoured Western ways of knowing, learning and teaching, and delimited opportunities for non-white students. (See, for a broader discussion on this movement, Jansen 2019).
 6. Madlingozi conceives of three major legacies of colonialism in Africa, of which 'a world of apartness' is one, together with 'the colonial state form, and conversely the eternal subjugation of indigenous sovereignties' and 'the continuing subordination of African life-worlds and their epistemologies and jurisprudences' (2018).
 7. See, too, Saini 2019.
 8. Following the publication of *Decolonising the Mind* in 1986, wa Thiong'o published only in his native African languages, Gikuyu and Swahili.
 9. For this reason, the work being done by data scientists in developing and promoting AI systems, such as machine learning and natural language processing, in African languages is so important. See Marivate et al. 2020 and Martinus and Abbott 2019.
 10. Hence, Aimé Césaire calls for a new humanism 'made to the measure of the world' (2001, 73).
 11. See also note 5.
 12. 'The redefining and re-signifying of life in conditions of dignity' (Mignolo and Walsh 2018, 3).
 13. In particular, their work on the 'gender shades' project which revealed the inordinate level of misrecognition of black female faces, in particular, by the facial recognition technologies of IBM (Buolamwini and Gebru 2018). See also, <http://gendershades.org/>.
 14. See also, Breckenridge (2014) who traces the rise of the biometric state form in South Africa with the use of statistical biometric tools to monitor the movements of the native population under colonialism and apartheid.
 15. See Appadurai on caste divisions in India, 1993.
 16. See also Sawyer Seminar on 'Histories of AI: A Genealogy of Power,' wherein the idea of data systems functioning as a palimpsest is discussed, <https://www.hps.cam.ac.uk/about/research-projects/histories-of-ai/activities/reading-group-graduate-training/rg1>.

17. That is, an advanced form of AI that can, like human intelligence, work at a generalized level, rather than simply perform specific, localised intelligence tasks, which is where the standard of the field currently lies.
18. See also note 5 above on the work of indigenous communities in developing a decolonial AI.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Rachel Adams is a researcher based in South Africa whose work explores critical perspectives of Artificial Intelligence and digital change in subaltern contexts.

ORCID

Rachel Adams  <http://orcid.org/0000-0003-1436-190X>

References

- Achiume, E. Tendayi. 2020. "Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis." Report of the Special Rapporteur on contemporary forms of racism, racial discrimination, xenophobia and related intolerance to the Forty-Fourth Session of the United Nations Human Rights Council, June 18. A/HRC/44/57.
- Adams, Rachel. 2019. "Helen 'A'Loy and Other Tales: A Gendered Reading of the Narratives of Hopes and Fears in Intelligent Machines and Artificial Intelligence." *AI and Society*.
- Adams, Rachel. 2020. *Transparency: New Trajectories in Law*. London and New York: Routledge.
- Adebisi, Foluke Ifejoba. 2019. "Why I say Decolonisation is Impossible." <https://folukeafrica.com/why-i-say-decolonisation-is-impossible/>.
- Appadurai, Arjun. 1993. "Number in the Colonial Imagination." In *Orientalism and the Postcolonial Predicament: Perspectives on South Asia*, edited by Carol A. Breckenridge and Peter van der Veer, 314–339. Philadelphia: University of Pennsylvania Press.
- Arun, Chinmayi. 2020. "AI and the Global South: Designing for Other Worlds." In *Oxford Handbook on AI Ethics*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das. Oxford: Oxford University Press.
- Ballim, Faeza, and Keith Breckenridge. 2018. "Divinatory Computation: Artificial Intelligence and the Future of the African Continent." Academic Paper, July 25.
- Baudrillard, Jean. 1994. *The Precession of Simulacra*. (Trans. Ann Arbor). Michigan: University of Michigan Press.
- Bell, Genevieve. 2018. "Decolonising AI." Lecture Delivered at the Australia National University.
- Benjamin, Ruha. 2019. *Race After Technology: The New Jim Crow*. Durham: Duke University Press.
- Biesheuvel, Simon. 1943. *African Intelligence*. Johannesburg: South African Institute of Race Relations.

- Biesheuvel, Simon. 1958. "Objectives and Methods of African Psychological Research." *The Journal of Social Psychology* 47: 161–168.
- Birhane, Abeba. 2019. "The Algorithmic Colonization of Africa." *Real Life*, July 18. <https://reallifemag.com/the-algorithmic-colonization-of-africa/>.
- Blackwell, Alan. 2021. "Ethnographic Artificial Intelligence." *Interdisciplinary Science Review*, this issue.
- Bonneuil, Christophe. 2000. "Development as Experiment: Science and State Building in Late Colonial and Postcolonial Africa, 1930–1970." *Osiris: Nature and Empire: Science and the Colonial Enterprise* 15: 258–281.
- Breckenridge, Keith. 2014. *Biometric State: The Global Politics of Identification and Surveillance in South Africa, 1850 to the Present*. Cambridge: Cambridge University Press.
- Browne, Simone. 2015. *Dark Matters: On the Surveillance of Blackness*. Durham, NC: Duke University Press.
- Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1–15.
- Cave, Stephen. 2020. "The Problem with Intelligence: Its Value-Laden History and the Future of AI." AIES '20: AAAI/ACM Conference on AI, Ethics, and Society Proceedings, New York, February 7–8. <https://doi.org/10.1145/3375627.3375813>.
- Césaire, Aimé. 2001. *Discourse on Colonialism* (Trans. Joan Pinkham). New York: Monthly Review Press.
- Chenweizu. 1987. *Decolonising the African Mind*. Lagos, Nigeria: Pero Press.
- Chinoy, Sahil. 2019. "The Racist History behind Facial Recognition." *New York Times*, July 10. <https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html>.
- Chollet, François. 2019. "On the Measure of Intelligence." *Computer Science*.
- Couldry, Nick, and Ulises A. Mejias. 2019. "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject." *Television and New Media* 20 (4): 336–349.
- Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, et al. *AI Now 2019 Report*. New York: AI Now Institute. https://ainowinstitute.org/AI_Now_2019_Report.html.
- Crawford, K., and V. Joler. 2018. "Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources." *AI Now Institute and Share Lab*. <https://anatomyof.ai>.
- Falk, Michael. 2021. "Artificial Stupidity." *Interdisciplinary Science Reviews*, this issue.
- Foucault, Michel. 1982. "The Subject and Power." In M Foucault *Power: Essential Works of Foucault 1954–1984 Volume 3* (trans. R Hurley & Others, 2000) 326.
- Frischmann, Brett, and Evan Selinger. 2018. *Re-Engineering Humanity*. Cambridge: Cambridge University Press.
- Garvey, Colin. 2019. "Artificial Intelligence and Japan's Fifth Generation: The Information Society, Neoliberalism, and Alternative Modernities." *Pacific Historical Review* 88 (4): 619–658. <https://doi.org/10.1525/phr.2019.88.4.619>.
- Gikanda, Simon. 2015. "Rethinking the Archive of Enslavement." *Early American Literature* 50 (1): 81–102.
- Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. 2019. "Better, Nicer, Cleaner, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." Proceedings of the 52nd Hawaii International Conference on Systems Science.
- Grosfoguel, Ramón. 2007. "The Epistemic Decolonial Turn." *Cultural Studies* 21 (2-3): 211–223.

- Grosfugal, Ramon. 2011. "Decolonizing Post-Colonial Studies and Paradigms of Political-Economy: Transmodernity, Decolonial Thinking, and Global Coloniality." *TRANSMODERNITY: Journal of Peripheral Cultural Production of the Luso-Hispanic World* 1 (1).
- Hogarth, Ian. 2018. "AI Nationalism." <https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism>.
- Holzmeyer, Cheryl. 2021. "Beyond "AI for Social Good" (AI4SG): Social Transformations - Not Tech-Fixes - for Health Equity." *Interdisciplinary Science Reviews*, this issue. doi:10.1080/03080188.2020.1840221.
- Jansen, Jonathan. (ed). 2019. *Decolonisation in Universities: The Politics of Knowledge*. Johannesburg: Wits University Press.
- Jansen, Jan, and Jurgen Osterhammel. 2017. *Decolonization: A Short History (Translated Jeremiah Riemer)*. Princeton, New Jersey: Princeton University Press.
- Jobin, Anna, Marcello Ienca, and Vayena, Effy. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1: 389–399.
- Kalluri, Pratyusha. 2020. "Don't Ask if Artificial Intelligence is Good or Fair, Ask How It Shifts Power." *Nature* 583: 169. <https://www.nature.com/articles/d41586-020-02003-2>.
- Kalpagam, U. 2000. "The Colonial State and Statistical Knowledge." *History of the Human Sciences* 13 (2): 37–55.
- Kalpagam, U. 2014. *Rule by Numbers: Governmentality in Colonial India*. London: Lexington Books.
- Keyes, Os. 2018. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." Proceedings of the ACM on Human-Computer Interaction, November, Article No.: 88. <https://doi.org/10.1145/3274357>.
- Keyes, Os, Zoe Hitzig, and Mwenza Blell. 2021. "Truth from the Machine: Artificial Intelligence and the Materialisation of Identity." *Interdisciplinary Science Reviews*, this issue.
- Laher, Sumaya, and Kate Cockcroft. 2014. "Psychological Assessment in Post-Apartheid South Africa: The Way Forward." *South African Journal of Psychology* 44: 303–314.
- Mabhena, Cetshwayo Zindabazwe. 2017. "From Decolonisation to Decolonialisation." *Sunday News*, January 29. <https://www.sundaynews.co.zw/from-decolonisation-to-decolonialisation/#:~:text=Decolonialisation%20therefore%2C%20is%20the%20process,killing%20the%20ghost%20of%20colonialism>.
- Madlingozi, Tshepo. 2018. "The Proposed Amendment to the South African Constitution: Finishing the Unfinished Business of Decolonisation?" *Critical Legal Thinking*, April 6. <https://criticallegalthinking.com/2018/04/06/the-proposed-amendment-to-the-south-african-constitution/>.
- Marivate, V., T. Sefara, V. Chabalala, K. Makhaya, T. Mokgonyane, et al. 2020. "Investigating an Approach for Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi." arXiv preprint arXiv:2003.04986.
- Martin, A. L. 1916. "Experiments with Binet-Simon Test upon African Colored Children." *The Training School Bulletin* 12: 142–143.
- Martinus, L., and J. Z. Abbott. 2019. "A Focus on Neural Machine Translation for African Languages." arXiv preprint arXiv:1906.05685.
- Mbembe, Achille. 2017. *Critique of Black Reason*. Johannesburg: Witwatersrand University Press.
- Mbembe, Achille. 2019. *Necropolitics*. Johannesburg: Witwatersrand University Press.
- Mhlambi, Sabelo. 2020. "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance." Carr Centre

- Discussion Paper. https://carrcenter.hks.harvard.edu/files/cchr/files/ccdp_2020-009_sabelo_b.pdf.
- Mignolo, Walter D., and Catherine E. Walsh. 2018. *On Decoloniality*. London: Duke University Press.
- Modiri, Joel. 2019. *Key Note Address: Decolonisation and the Law School*. University of Bristol.
- Mohamed, Shakir. 2018. "Decolonising Artificial Intelligence." <http://blog.shakirm.com/2018/10/decolonising-artificial-intelligence/>.
- Mohamed, Shakir, Marie-Therese Png, and William Isaac. 2020. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *Philosophy and Technology*.
- Ndlovu-Gatsheni, Sabelo. 2015. "Decoloniality as the Future of Africa." *History Compass* 13 (10): 485–496.
- Ndlovu-Gatsheni, Sabelo. 2017. "Perhaps Decoloniality is the Answer? Critical Reflections on Development from a Decolonial Epistemic Perspective." *Africanus: Journal of Development Studies* 43 (2): 1–11.
- Ndlovu-Gatsheni, Sabelo. 2020. "Decolonization, Decoloniality, and the Future of African Studies: A Conversation with Dr. Sabelo Ndlovu-Gatsheni." *Social Sciences Research Council Blog*, January 15. <https://items.ssrc.org/from-our-programs/decolonization-decoloniality-and-the-future-of-african-studies-a-conversation-with-dr-sabelo-ndlovu-gatsheni/>.
- Ni Loideain, Nora, and Rachel Adams. 2019. "From Ava to Siri: The Gendering of Virtual Personal Assistants and the Role of Data Protection Law." *Computer Law and Security Review*.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.
- Peña, Paz, and Joanna Varon. 2019. "Decolonising AI: A Transfeminist Approach to Data and Social Justice." GIS Watch 2019: Artificial Intelligence: Human Rights, Social Justice and Development, Association for Progressive Communication, Article 19 and Swedish International Development Cooperation Agency.
- Qin, Hongwei, Junjie Yan, Xiu Li, and Xiaolin Hu. 2016. "Joint Training of Cascaded CNN for Face Detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3456–3465.
- Quijano, Anibal. 2010. "Coloniality of Power, Eurocentrism and Latin America." *Nepantla: Views from South* 1 (3): 533–580.
- Quijano, Anibal. 2017. "'Good Living': Between Development and the De/Coloniality of Power." In *The Routledge Companion to Inter-American Studies*, edited by Wilfried Raussert, 363–371. London and New York: Routledge.
- Ricaurte, Paola. 2019. "Data Epistemologies, The Coloniality of Power, and Resistance." *Television and New Media* 20 (4): 350–365.
- Said, Edward. 1978. *Orientalism*. New York: Vintage Books.
- Saini, Angela. 2019. *Superior: The Return of Race Science*. Boston, MA: Beacon Press.
- Sehlapelo, Martin, and Martin Terre Blanche. 1996. "Psychometric Testing in South Africa: View from Above and Below." *Psychology in Society* 21: 49–59.
- Smuts, Jan. 1930. "African Settlement. Part 1." *Journal of the Royal African Society* 29 (114): 109–131.
- Soudien, Crain. (n.d. unpublished manuscript). *A Short History of Race*.
- Spivak, Gayatri Chakravorty. 1988. "Can the Subaltern Speak?" In *Marxism and the Interpretation of Culture*, edited by C. Nelson, and L. Grossberg, 271–315. Urbana: University of Illinois Press.
- Taylor, Jane. 2019. "PAN: A Performance Lecture." *Critical Times* 2 (3): 493–517.

- Tilley, Helen. 2011. *Africa as a Living Laboratory: Empire, Development, and the Problem of Scientific Knowledge, 1870-1950*. Chicago: University of Chicago Press.
- Tuck, Eve, and K. Wayne Yang. 2012. "Decolonization is not a Metaphor." *Decolonization: Indigeneity, Education & Society* 1 (1): 1–40.
- Ulnicane, Inga, Damian Okaibedi Eke, William Knight, George Ogoh, and Bernd Carsten Stahl. 2020. "Good Governance as a Response to Discontents? Déjà vu or Lessons for AI from Other Emerging Technologies." *Interdisciplinary Science Reviews*, this issue.
- USAID. 2019. "Breaking News: USAID launches \$4.1 Million Artificial Intelligence Initiative." *ICT Works*, April 1. <https://www.ictworks.org/usaid-artificial-intelligence-initiative/#.XxmPVG5uK1M>.
- Vincent, James. 2017. "Putin says the nation that leads in AI 'will be the ruler of the world'". *The Verge*. <https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world>.
- Wa Thiong'o, Ngũgĩ. 1986. *Decolonising the Mind: The Politics of Language in African Literature*. New Hampshire: Heinemann Educational.
- Weinberg, Lindsay. 2019. "The Rationalization of Leisure: Marxist Feminism and the Fantasy of Machine Subordination." *Lateral* 8 (1).
- Williams, James. 2018. *Stand Out of Our Light: Freedom and Persuasion in the Attention Economy*. Cambridge: Cambridge University Press.
- Zuboff, Shoshana. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Profile Book.



Ethnographic artificial intelligence

Alan F. Blackwell

Professor of Interdisciplinary Design, Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

ABSTRACT

Artificial intelligence research is seldom conducted ethnographically. Nevertheless, the field itself has a distinctive culture – an engineering imaginary within which ‘general intelligence’ is considered to be a phenomenon that can be both defined and tested by reference to some universal (probably mathematical) standard. This paper describes a research agenda that sets out to question those assumptions, through a programme of ethnographic field work, collaborating with computer scientists and educators in several countries of sub-Saharan Africa.

KEYWORDS

Artificial intelligence; critical technical practice; human-computer interaction; cognitive variations; broad learning

Introduction

This article sets out an alternative course of enquiry for a discontented AI engineer, in which I have stepped away from the usual concerns of my field to ask more searchingly where the real problems are. The problems I address, while often apparent to critics of AI, are not necessarily salient concerns for engineers themselves. In my own case, new priorities came from simply following an engineering logic beyond the expected boundaries of the discipline, initially through the commercial concerns of designing more marketable ‘user-centric’ software products, and subsequently, through collaboration with colleagues in the Global South, as Director of an interdisciplinary initiative advancing research for the Sustainable Development Goals (SDGs).

After training as an artificial intelligence engineer in the 1980s, my professional work applying AI methods in industrial and commercial settings led me to the conviction that in most projects, it was human questions rather than technical ones that turned out to be most problematic. This increasing interest in ‘human factors’ (in traditional engineering) or ‘human–computer interaction’ (HCI, as it is called in computer science) eventually drew me away from engineering to start a mid-career PhD at a centre for cognitive neuroscience.

CONTACT Alan F. Blackwell  afb21@cam.ac.uk  University of Cambridge, Department of Computer Science and Technology, William Gates Building 15 J J, Thomson Avenue Cambridge, CB3 0FD, United Kingdom

© 2020 Institute of Materials, Minerals and Mining Published by Taylor & Francis on behalf of the Institute

However, this initial step away from engineering was more disconcerting than I expected. The problem was not that I knew too little about my new discipline, but apparently knew too much. I had always considered that the AI engineering methods I used were related to human cognition only metaphorically, having an engineer's typical scepticism for the 'strong AI' agenda. I was surprised to find on arrival in a psychology department, that while I had never studied the subject, on the basis of my earlier engineering work I was considered well qualified to understand the human mind. I learned that terms I had considered figurative in relation to system modules such as 'learning', 'memory' and 'recognition', were treated in cognitive science as literal components of the mind.

Twenty years later, I understand that my surprise resulted from the habit described by Philip Agre (1997) as 'AI's elastic use of language'. I had previously been aware of Agre as a well-informed commentator on technology, but not for his critique of AI as unacknowledged philosophy. An interest in philosophy was indeed what had attracted me to AI (from reading philosophical texts as a hobby during my engineering degree). However, if I had read Agre's critical work before starting my PhD, I might have been less surprised by the overly broad claims of cognitive science. I might also have been in a position to define a methodology for my research instead of, as Agre observed of his own critical development during the period that I myself was studying AI, lacking a methodology because our field lacked clear critical consciousness and purpose.

Instead, my explorations have been somewhat haphazard. An earlier volume of this journal explored wide-ranging questions arising from Geoffrey Lloyd's historical and anthropological work on *Cognitive Variations* (2007). My discussions with Lloyd had led me to question one of the more quixotic applications of AI methods: the universal standardization of human emotion. In my contribution to that issue, I traced the origins of the 'affective computing' enterprise to the engineering impulse for systematization, fuelled by fantasies of escape from the complexities and obligations of emotional reciprocity, where a person fascinated by technology aspires to become a man-machine (Blackwell 2010).

A second set of concerns came about from questions of ethics in AI research. I had chaired a working group that established ethical review procedures for the School of Technology in Cambridge University. But I was surprised, as the ethics of AI research came to be debated in public forums, that these debates apparently paid no attention to the day-to-day ethical problems that were encountered by my own research review board, instead concentrating on distant speculations such as the emergence of an apocalyptic 'Singularity' (e.g. Kurzweil 2005; following Vinge 1993) or pursuing libertarian preoccupations such as freedom from government surveillance and social obligation.

A third set of concerns came about through my work as director of Cambridge Global Challenges, a university-wide strategic research initiative prioritizing research that can advance the SDGs in the low-income countries. I had already been aware that much AI functionality depended on the unseen labour now described as ‘ghost work’, and that the economics of such labour was particularly exploitative of those who are already disadvantaged (Blackwell 2019a, 2019b). But as other contributors to this issue (e.g. Ulnicane et al.; Adams) have noted, this patterned inequality is compounded by the many ethical conventions for AI that have been drafted, which include no voices from the Global South.

These haphazard developments of my own discontent, as with the formative experiences reported by Philip Agre, are hardly comprehensive. Furthermore, I have continued to learn (since drafting the original version of this paper), that pulling on these disparate loose threads in the fabric of sociotechnical AI does not make me well-equipped to understand the systemic problems of colonialism, or the ways in which the legacy of scientific racism has constructed so many of the foundational principles of AI and cognitive science. I return to discuss those questions in the last part of this article, although the work by Adams and Keyes, Hitzig, and Blell in this issue offers a far more extensive consideration of these important questions.

The rest of this paper sets out an agenda that comes from exploring these concerns from within the technical establishment, using the ethnographic methods of HCI research, together with the historical/sociological perspective that I have previously used to pursue internal reflections on my discipline (e.g. Blackwell 2006; Blackwell, Blythe, and Kaye 2017). In what follows, I first propose an alternative critical perspective from which to think about AI in relation to historical cultural practices of the African continent, rather than the European one. Second, I explore the dynamics through which AI is performed in culturally situated ways. Third, I explain the ethnographic research agenda motivated by this understanding of AI as culturally situated. And finally, as often advocated in HCI, I set out some research questions that might lead to ‘implications for design’ (Dourish 2006) – ways that AI could be imagined differently by its engineers, and above all by engineers in Africa, engineers of the Global South, indigenous engineers, and any other engineers who question the assumptions of privilege that are embedded within global knowledge technosystems inherited from racist colonialism.

Historical performances of AI in Africa and Europe

In his classic narrative of Igbo culture *Things Fall Apart*, Chinua Achebe describes a village ceremony at which Evil Forest and other *egwugwu* spirits¹ emerge from their sacred house:

And then the *egwugwu* appeared. The women and children sent up a great shout and took to their heels. It was instinctive. A woman fled as soon as an *egwugwu* came in sight. And when, as on that day, nine of the greatest masked spirits in the clan came out together it was a terrifying spectacle. [...] Each of the nine *egwugwu* represented a village of the clan. Their leader was called Evil Forest. [...]

Okonkwo's wives, and perhaps other women as well, might have noticed that the second *egwugwu* had the springy walk of Okonkwo. And they might also have noticed that Okonkwo was not among the titled men and elders who sat behind the row of *egwugwu*. But if they thought these things they kept them within themselves. The *egwugwu* with the springy walk was one of the dead fathers of the clan. He looked terrible with the smoked raffia body, a huge wooden face painted white except for the round hollow eyes and the charred teeth that were as big as a man's fingers. On his head were two powerful horns.

When all the *egwugwu* had sat down and the sound of the many tiny bells and rattles on their bodies had subsided, Evil Forest addressed the two groups of people facing them. (Achebe 1958/2010, 84–86)

The European continent has contrasting traditions of elaborately costumed oracle. In 1770, Wolfgang von Kempelen demonstrated his chess-playing automaton 'the Turk' to Empress Maria Theresa of Austria, prompting public fascination that would continue for over 50 years of exhibitions and performances throughout Europe and America. Built to hide an expert chess-player concealed within, the Turk was a mannequin dressed in oriental costume, seated behind a fake clockwork mechanism whose purpose was to misdirect the audience (Figure 1).

Where the appearance of the Igbo *egwugwu* was designed to evoke the presence and spirit of ancestors, automata such as the Turk evoke the mechanical ideals and political economy of the Age of Reason (Schaffer 1999). The skill of the mechanic/inventor is demonstrated not only in the construction of the clockwork apparatus, but also in the stage performance when he opens doors in the cabinet with a flourish, allowing the audience to see past the clockwork through an otherwise empty box.²

The modern performance of AI

These distinctive costumed/puppet performances from the history of Western Europe and Western Africa can both be interpreted as variants on the modern 'science' of artificial intelligence (AI). In AI, as with both the *egwugwu* and the Turk, the outward appearance demonstrates skilled material construction,

¹I am indebted to Rachel Adams for suggesting the comparison of artificial intelligence to *egwugwu*, which is to be developed further in her own forthcoming book.

²As each door is opened in turn, the hidden chess-player slides to the other side, in the same manner as when stage magicians demonstrate that they have sawn an assistant in half.

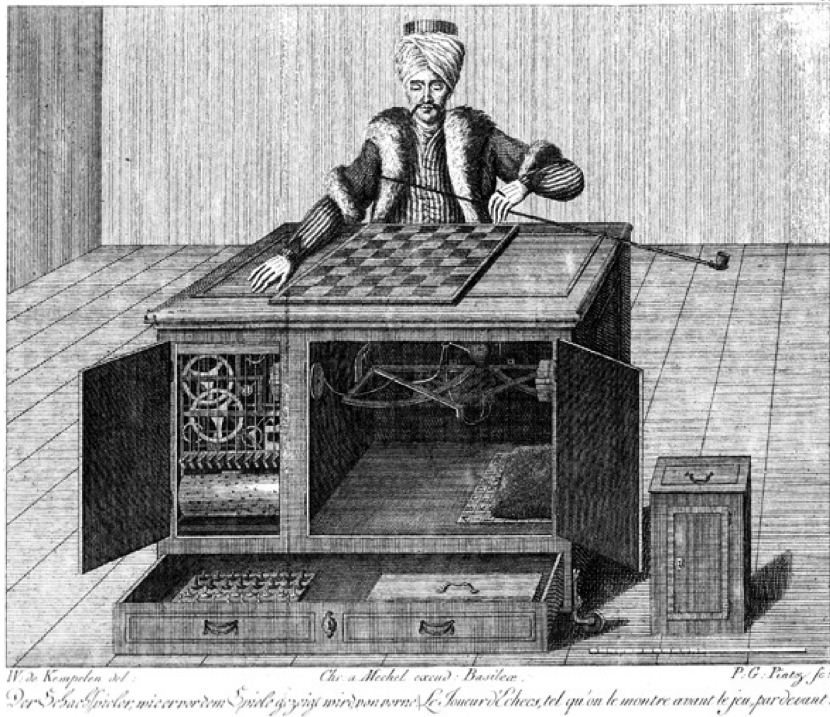


Figure 1. Copper engraving from Karl Gottlieb von Windisch, *Briefe über den Schachspieler des Hrn. von Kempelen, nebst drei Kupferstichen die diese berühmte Maschine vorstellen*. 1783. Source: Wikimedia Commons, public domain licence.

while the behaviour of the magical puppet is expected to demonstrate a combination of judgment, prognostication, and intellectual skill.

In our new digital age of reason, computational infrastructure underlies oracular performances. Global corporations such as Amazon provide the infrastructural components of AI systems, ranging from the supply of cloud computation via Amazon Web Services, to the commissioning of ‘human intelligence tasks’ (HITs) through the labour market of Amazon Mechanical Turk. This new Mechanical Turk was named in direct homage to the puppet of von Kempelen’s stage apparatus, alluding to the fact that AI researchers also need to embed (and potentially conceal) real human expertise inside their demonstrations of technical skill (Irani and Silberman 2013).

I placed the word ‘science’ in quotation marks, in the introduction to this section, because when analysing the origins and purposes of its public performances, AI might be considered a branch of literature as much as science. The day-to-day occupation of AI researchers is to imagine some behaviour that might be exhibited by their systems (for example, a behaviour that would serve to demonstrate properties that might be described as ‘learning’, ‘conversation’, ‘perception’, or ‘consciousness’), and then to create algorithms that will

produce this imagined behaviour. Algorithms are constructed as texts, composed in a programming language. Although programmes are relatively formalized texts, resembling film scripts or plays more closely than prose (Laurel 1993), they are still undoubtedly literary works (Cox and McLean 2013), protected by copyright, and attributed to one or more authors. These essential elements of textuality and imagination underlie one popular definition of AI, as ‘the branch of computer science that is devoted to making computers work the way they do in the movies’.

Other definitions of AI also draw on traditions of performance and illusion, as for example demonstrated by the perennial interest in the Turing Test. Alan Turing originally proposed his Imitation Game as a positivist strategy to circumvent the potentially intractable philosophical problem of how one should define ‘intelligence’. Turing suggested that if we wish to test whether a machine is intelligent, we do not need further definition, but can simply compare the outward behaviour of two potentially intelligent actors. Since we acknowledge humans to be intelligent, then if the observed behaviour of a machine cannot be distinguished from a human, this offers a falsifiable test of equivalence (see also Collins, this issue).

Unfortunately, Turing’s Imitation Game (as the name implies) can be interpreted as an exercise in illusion rather than one of science. Von Kempelen’s Turk imitated the skilled performance of a chess player, and thus appeared to demonstrate that a machine (although actually the person inside it) has those same intelligent skills. The construction of this illusion, as with all magic, depends on a careful management of the expectations of the observers, including essential constraints on what they are allowed to see or ask. The game of chess, whose internal rules do not depend on social context, is a clever choice for a demonstration where the impostor must stay hidden inside a box. Similarly, the *egwugwu* are constrained to engage only in those activities appropriate to ancestral spirits (they do not, for example, share houses with ordinary people), with the result that the illusion is not disrupted by any action that might reveal what is inside their costumes.

The prestigious Loebner Prize,³ awarded annually to software systems that come closest to passing the Turing Test, is generally won by systems that most ingeniously exploit such contextual constraints. A famous example was a chatbot that assumed the persona and emulated the conversational style of an uncooperative teenager, apparently convincingly enough to persuade several judges of the contest. The success of specialist AI demonstrators, for example, winning games of chess, of Go and so on, can similarly be attributed to the very constrained context and rules of such games.⁴

³<https://www.aisb.org.uk/events/loebner-prize>

⁴The social implications and limitations of such tests are explored with far greater detail and sophistication by Harry Collins (2018).

Just as these public demonstrations of the science of AI rely on careful management of the performance constraints, so the commercial successes of AI can also be attributed to the successful definition of system constraints, defining a broader problem in sufficiently narrow terms that it can be solved by a specialist algorithm. As AI researchers construct demonstrations that are intended to exhibit ‘perception’, ‘learning’, and so on, these demonstrations do often turn out to have some industrial application – for example, experiments in perception have led to successful industrial applications of computer vision, while experiments in learning have resulted in a current boom in the application of statistical inference and data science through machine learning methods.

For commentators on AI, a critical distinction is often made between these kinds of specialist system (in which the exercise of imagination eventually results in reliable but relatively mundane technical functionality), and the grander ambition of ‘artificial general intelligence’ (AGI), which is anticipated to replicate human performance more completely, and would no longer be constrained to special contexts or applications.

Of course, this scientific ambition might become problematic, if it were the case that *all* demonstrations of AI, like the *egwugwu* or von Kempelen’s Turk, rely on carefully managed performances to obscure human agents concealed within (Blackwell 2019b). Furthermore, the ambition of a uniform and mechanizable ‘general intelligence’, when juxtaposed with the *egwugwu* and the Turk, seems to ignore the likelihood of cognitive variations when intelligence is studied in other contexts (Lloyd 2007). Attempts to circumvent all these problems, as in the recent comprehensive proposal by Chollet (2019) for a universal benchmark that would measure the performance of AGI in ways that are both practically relevant and also culturally non-specific, appear to founder on their internal contradictions. Chollet’s proposal explicitly rejects any need for the AGI to have a body, leaving future competitors able to demonstrate their puppets through staged performances that would be even more constrained than those of the *egwugwu* and the Turk. Just as with the attempt to create ‘culture-fair’ IQ tests, such efforts at metrickation seem inevitably tainted with culture. Indeed, as demonstrated by Cave (2020), the concept of intelligence itself is apparently indelibly tainted as an exercise of cultural hegemony and racism. These concerns, when taken together, demand completely fresh approaches to the technical study of AI.

Inventing AI elsewhere

The questions raised in this discussion form the motivation for a programme of fieldwork to explore the question, What would AI look like if it were invented in Africa? Even to ask this question might appear provocative or insensitive, if it were interpreted as an attempt to discover a primordial kind of AI, or to reinforce neo-colonial economics by offering cheaper and less powerful

versions of Western technologies to low-income countries. If AI is a science, why should science be any different in Africa? But if AI is literature, then how could it possibly be the same in Africa? I am guided here by the insights of Helen Verran, whose experiences of working with mathematics teachers in Nigeria led her to observe that:

I stand as a participant caught up in complex power relations. Being both teacher and learner and at the same time a theorist, a storyteller of the episodes, reciprocal indebtedness and accountability characterize my position. [...] The explanations of the 'problem' of science and mathematics curricula in places like Ife-Ife that grow from these accounts, and the solutions that follow from those explanations, remake a colonizing modernity [...] Foundationist explanations fail the critical project. They actually make it impossible to imagine futures different from pasts. [...] Keeping the delight and curiosity potent is important. I want to privilege the disconcertment. This involves imagining a way of doing critique that effects neither the dominating unity of universalism, nor the oppositional fragmentation of relativism. (Verran 2001, 36)

Like Verran, my own goal is to 'privilege disconcertment', as in Bidwell's reasoning from Verran to question commensurability between database schemas and Xhosa persons (Bidwell 2016). To the extent that AI research is a literary and cultural endeavour, addressing the appetites and concerns of particular audiences, it seems clear not only that *egwugwu* in Nigeria might be constituted differently from the Turk in eighteenth century Austria, but that the science fiction imagination of contemporary writers in the African continent might address different issues from the science fiction imagination of the USA or UK. One goal of this fieldwork is to become more closely engaged with the exercise of engineering imagination among technologists and commentators who live in African countries.

When working in low-income countries, the AI engineer must also address the fact that these technologies represent extreme dynamics in the policies and infrastructure of digital capitalism (Crawford and Joler 2018). The labour markets of Amazon Mechanical Turk, the 'user generated content' of Facebook and Twitter, and the surveillance capitalism of Google result in 'giant automated plagiarism machines' (Poole 2019), or 'subjectivity factories' that collect data from users in order to then sell it as the performances of AI agents (Blackwell 2019b). The dynamics of human work, financial reward, and media consumption that underlie this infrastructure are constructed and negotiated solely by privileged members of the elite within wealthy countries (Zuboff 2019).

It is very likely that mundane intellectual tasks could also be automated, or information transferred more effectively, in African countries – a change anticipated by some of those countries as a potential 'Fourth Industrial Revolution' that may perhaps gain them some advantage in the new world order. But there is little evidence that the existing infrastructure of AI has been built to

benefit the people of those countries, or to expect that it will be built in such a way in future. A second goal of this fieldwork is therefore to better understand the kinds of cognitive labour that are valued in the Global South, in order to identify sociotechnical economic principles on which it can be equitably enhanced.

A third goal of this fieldwork is to consider how the standard tools and practices of AI research might be modified to become more effective in the hands of African AI researchers. Much current research is conducted using software tools (programming languages, data sets, and computation architectures) that have been developed in corporate laboratories with industry investment. Many of these tools are made ‘open’ for use by academic researchers, though many suspect that free distribution serves primarily to gain and reinforce commercial advantage through widespread adoption of one company’s standards and infrastructure rather than another.

It is notable that many of these standard tools share an emphasis on ‘Big Data’, ‘deep learning’, and deployment of resources so expensive that they are beyond the reach of many Western academics—let alone those in the Global South. But there is no *a priori* reason to believe that AI must always be associated with profligate corporations, any more than we would expect all people in the nineteenth century to prefer the trappings of a mechanical chess player rather than an Igbo ancestral spirit. This third goal will involve the development of new concepts for programming languages (as well as the educational resources and software tools associated with such languages) that can be used to express distinctively African conceptions of intelligence, taking into account the imaginative and economic motivations arising from explorations of the first and second goals. I refer to this alternative emphasis as the pursuit of ‘broad learning’, rather than ‘deep learning’.

This final part of my research agenda is oriented toward technological change, and indeed represents a form of action research, rather than purely ethnographic enquiry. As with all fieldwork, it depends fundamentally on the specifics of particular situations, rather than scientific abstractions such as AGI, with their problematic legacies of racism and misogyny. Devlin and Belton, drawing on N. Katherine Hales, observe that the perspective through which the rational mind is separated from a body, ‘could only come from the privileged group of white, able-bodied men: most outside that group are keenly aware that the ways their bodies are read by society impacts their day-to-day existence as a marginalized subject’ (2020).

Discussion – some implications for design?

The concerns of this paper were introduced by considering how AI, as commonly presented, is a cultural artefact. AI is shaped by culturally-specific imaginaries and implemented through the situated actions of cultural agents,

including the engineers that create algorithms, the workers ‘hidden’ within the elaborate technological costumes, and businesspeople who deploy rhetoric and showmanship to package and sell the resulting spectacle as ‘AI’.

By contrast, in HCI research it is well understood that cultural and economic relations can be explored more effectively through ethnographic methods, which are routinely adapted and taught to engineers as a basis for system design (e.g. Holtzblatt and Beyer 1997). However, the fieldwork that I am engaged in moves beyond those framing assumptions to offer the privilege of learning alternative frames and perspectives from people who may be less habituated than I am to the technological narratives of contemporary digital capitalism. I also hope that my own professional experiences as an AI engineer and educator over several decades will offer some reciprocal benefit when working in countries that have fewer resources to invest in digital infrastructures.

Apart from generally critical alertness to fallacies in current popular thinking and policy, and an ethnographic determination to look and listen, is there any further theoretical stance that might inform design opportunities in this project?

One possibility is the cross-cultural investigation of mechanized know-how in the digital era, rejecting the notion of AGI as formalized by researchers such as Chollet (2019), as well as the colonial and racist conceptions of ‘intelligence’ itself as revealed by Adams (this issue) and Cave (2020), to instead pursue long-standing metaphysical questions in epistemology and theory of mind, in the same manner as Helen Verran’s (2001) investigations of different logics of science. Philip Agre, in his call for a critical technical practice of AI (1997), concluded with the observation that AI as it proceeds in reality is a craft practice rather than a laboratory science. This is surely an opportunity to explore the boundaries of knowledge and materiality in ways other than those presumed by the construction, ‘artificial intelligence’.

A second is to contribute to an alternative economic basis that more equitably rewards the ‘work’ of attending to and instructing digital systems, for example, by advancing the agenda of Postcolonial Computing proposed by Philip, Irani, and Dourish (2012), or implementing the strategies proposed by Mohamed, Png, and Isaac (2020) toward ‘Decolonial AI’ (a problematic phrasing itself requiring more thorough investigation; see Rachel Adams, this issue). While this work has already drawn attention to the role of Amazon Mechanical Turk in maintaining labour structures (e.g. Irani 2015), an alternative theoretical perspective might address policy concerns about the ‘Fourth Industrial Revolution’ by revisiting the economics of Charles Babbage (1832), whose measurements of the division between skilled and unskilled labour in nineteenth century manufacturing both foreshadowed Taylorist automation and drew attention to the critical problems of integrating people into technological systems, as Marx recognized (Schaffer 1994). The rise of surveillance capitalism (Zuboff 2019) has been documented in relation to the wealthy

economies of the Global North, but it is as yet unclear how the micro-tasks of cognitive labour – clicks, likes, tweets, and the other everyday housekeeping tasks of the social media economy – might become instruments of extraction in poor countries. Some human intelligence tasks are already commoditized by Amazon Mechanical Turk, but this actuality does not by any means delimit future possibilities, for each and every interaction with digital systems can be directly quantified as seconds of a human lifespan, thereby abstracting conscious human attention itself into an economic commodity (Blackwell 2019a).

A third possible resource for conceptual design, perhaps of more specific interest to those who investigate science from a trans-cultural and trans-historical perspective, as in Verran's work with Nigerian teachers, is to re-examine the intellectual foundations of contemporary AI research. Here I have the privilege to draw on the work of my late friend David MacKay, whose early work applying principles from Claude Shannon's information theory and Thomas Bayes' mathematics of probability to experimental neural networks (MacKay 1992, 2003) was extraordinarily influential on the current generation of AI researchers.

Shannon's information theory casts doubt on the frequent claim made by boosters of extractive information technology that 'data is the new oil', or that 'Big Data' offers solutions to social and economic problems of all kinds. Shannon observed that, when data flows through any mechanical communication channel, the value it offers to the receiver depends not on the *quantity* of data, but the degree to which that data is novel or *surprising*. Because data that one has seen before is not surprising when it is received again, it has no value. Shannon's quantitative theory of information is thus a *theory of surprise*. Data that is surprising to one person (thus conveying a large amount of information) may not be at all surprising to another (thus conveying very little). It follows that quantitative measures of information cannot be purely objective, because the degree of surprise depends upon the contingent history and subjective state of the person(s) receiving it.⁵

Similar principles of subjective observation underlie Bayesian statistics, now in vogue amongst many contemporary AI researchers. Bayes explained that, when a person (for example, someone playing a game of chance) observes some event, this observation modifies their prior expectation about the likelihood of that event occurring. Therefore, Bayes' theorem can also be understood as a quantification of *surprise*: it expresses the relationship between a prior expectation and the new understanding that is gained as a result of a more or less surprising observation. The quantification of the prior likelihood can be understood as the amount of information one had before observing the event, and the quantified observation as the amount of new information that has been gained.

⁵In addition to the question of surprise, it is necessary to remember that one person's signal is another person's noise. Noise is surprising, but not valuable unless it can be related to what we know.

The foundations of Bayesian statistics were formulated over 150 years before the principles of conventional statistics that are now taught in pre-University mathematics courses. The older Bayesian statistics is not presently taught in schools, but does underlie much contemporary AI research.⁶ One may wonder how this situation has arisen. Might it be due to the positivist legacy of hypothesis testing that underlies much of the rhetoric of modern science? The foundations of conventional ‘frequentist’ statistics rely on presumption that the purpose of collecting data is to test hypotheses. The process by which the hypothesis is formed (on the basis of prior expectation) is explicitly excluded from consideration – and in fact made irrelevant through classroom teaching based on dice or coin tosses, devices explicitly designed to prevent anticipation of their outcomes. Consider by contrast the work of Jenny Gage and David Spiegelhalter (2018), who, drawing on their own experiences of teaching probability in schools and universities in Africa, have invented teaching devices and strategies that correspond more closely to practical judgments of likelihood.

In relation to the more advanced mathematics of computer science research, we might note that the Turing Test aims to define intelligence itself in a positivist experiment, and that the award of the Loebner Prize is specified in terms of a simple statistic (the proportion of the judging panel who are fooled by the candidate entries).

An alternative mode of description for AI might therefore be to explore information-theoretic concepts such as attention, expectation, likelihood, evidence, and observation through the ways that they are interpreted in other cultural contexts – including their implications for labour, attribution, agency, and the imagination of engineers, entrepreneurs, managers, policy-makers, and other actors involved in the social construction of information processing systems. An ethnographic account along these lines might offer opportunities for new digital tools and educational resources that would allow other societies to define and construct ‘artificial intelligence’ in, and on, their own terms.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Alan F. Blackwell is Professor of Interdisciplinary Design at the University of Cambridge, with degrees in engineering, computer science and psychology. He founded the Crucible Network for Research in Interdisciplinary Design, and is currently director of Cambridge

⁶And indeed, some earlier AI research, including the pioneering work of David MacKay’s father Donald MacKay, who offered a stringent critique of the ‘merchants of automata’ in his 1954 address to the International Congress of Psychology (MacKay 1956).

Global Challenges, a Strategic Research Initiative of the University of Cambridge. He is a Fellow of Darwin College, and a Fellow of the Royal Society of Arts.

References

- Achebe, C. 1958/2010. *Things Fall Apart*, Penguin Classics Edition. London: Penguin, 84–86.
- Agre, P. 1997. "Toward a Critical Technical Practice." In *Bridging the Great Divide: Social Science, Technical Systems, and Cooperative Work*, edited by G. Bowker, L. Gasser, S. L. Star, and B. Turner, 131–158. Mahwah, NJ: Lawrence Erlbaum.
- Babbage, C. 1832. *On the Economy of Machinery and Manufactures*. London: Charles Knight.
- Bidwell, N. J. 2016. "Moving the Centre to Design Social Media in Rural Africa." *AI & Society* 31 (1): 51–77.
- Blackwell, A. F. 2006. "The Reification of Metaphor as a Design Tool." *ACM Transactions on Computer-Human Interaction (TOCHI)* 13 (4): 490–530.
- Blackwell, A. F. 2010. "When Systemizers Meet Empathizers: Universalism and the Prosthetic Imagination." *Interdisciplinary Science Reviews* 35 (3–4): 387–403.
- Blackwell, A. F. 2019a. "Artificial Intelligence and the Abstraction of Cognitive Labour." In *Marx200: The Significance of Marxism in the 21st Century*, edited by M. Davis, 59–68. London: Praxis Press.
- Blackwell, A. F. 2019b. "Objective Functions: (In)Humanity and Inequity in Artificial Intelligence." *HAU: Journal of Ethnographic Theory* 9 (1): 137–146.
- Blackwell, A. F., M. Blythe, and J. Kaye. 2017. "Undisciplined Disciples: Everything You Always Wanted to Know About Ethnomethodology but Were Afraid to ask Yoda." *Personal and Ubiquitous Computing* 21 (3): 571–592.
- Cave, S. 2020. "The Problem with Intelligence: Its Value-Laden History and the Future of AI." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES'20)*, 29–35.
- Chollet, F. 2019. *On the Measure of Intelligence*. ArXiv:1911.01547 [Cs].
- Collins, H. 2018. *Artificial Intelligence: Against Humanity's Surrender to Computers*. Cambridge: Polity Press.
- Cox, G., and C. A. McLean. 2013. *Speaking Code: Coding as Aesthetic and Political Expression*. Cambridge, MA: MIT Press.
- Crawford, K., and V. Joler. 2018. *Anatomy of an AI System-The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources*. New York: AI Now Institute and Share Lab.
- Devlin, K., and O. Belton. 2020. "The Measure of a Woman: Fembots, Fact and Fiction." In *AI Narratives*, edited by S. Cave, K. Dihal, and S. Dillon, 357–381. Oxford: Oxford University Press.
- Dourish, P. 2006. "Implications for design." In *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI'06)* 541–550.
- Gage, J., and D. Spiegelhalter. 2018. *Teaching Probability*. Cambridge: Cambridge University Press.
- Holtzblatt, K., and H. Beyer. 1997. *Contextual Design: Defining Customer-Centered Systems*. San Francisco: Morgan Kaufmann.
- Irani, L. 2015. "The Cultural Work of Microwork." *New Media & Society* 17 (5): 720–739.
- Irani, L. C., and M. Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 611–620.

- Kurzweil, Ray. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York, NY: Viking.
- Laurel, B. 1993. *Computers as Theatre*. Boston, MA: Addison-Wesley.
- Lloyd, G. E. R. 2007. *Cognitive Variations: Reflections on the Unity and Diversity of the Human Mind*. Oxford: Oxford University Press.
- MacKay, D. M. 1956. "Towards an Information-Flow Model of Human Behaviour." *British Journal of Psychology* 47 (1): 30–43.
- MacKay, D. J. C. 1992. "Bayesian Interpolation." *Neural Computation* 4 (3): 415–447.
- MacKay, D. J. C. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.
- Mohamed, S., M. T. Png, and W. Isaac. 2020. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *Philosophy & Technology*. doi:10.1007/s13347-020-00405-8
- Philip, K., L. Irani, and P. Dourish. 2012. "Postcolonial Computing: A Tactical Survey." *Science, Technology, & Human Values* 37 (1): 3–29.
- Poole, S. 2019. "Deepfake or Fortune." *Guardian Review* 23 March 2019: 36–37.
- Schaffer, S. 1994. "Babbage's Intelligence: Calculating Engines and the Factory System." *Critical Inquiry* 21 (1): 203–227.
- Schaffer, S. 1999. "Enlightened Automata." In *The Sciences in Enlightened Europe*, edited by William Clark, Jan Golinski, and Simon Schaffer, 126–166. Chicago: University of Chicago Press.
- Verran, H. 2001. *Science and an African Logic*. Chicago: University of Chicago Press. 36.
- Vinge, V. 1993. "Technological Singularity." In *VISION-2: Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute*. 30–31.
- Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

