Truth from the Machine: Artificial Intelligence and the Materialisation of Identity

Os Keyes,[a] Zoë Hitzig[b] and Mwenza Blell[c]

[a]Department of Human Centered Design & Engineering, University of Washington, Seattle, WA, United States;
[b]Edmond J. Safra Center for Ethics, Harvard University, Cambridge, MA, United States;
[c]Policy, Ethics and Life Sciences Research Centre, School of Geography, Politics and Sociology, Newcastle University, Newcastle-upon-Tyne, United Kingdom

ARTICLE HISTORY
Compiled October 17, 2020

KEYWORDS
artificial intelligence; materialisation; epistemology; reification; naturalisation; race; sexuality; disability

ABSTRACT: In this paper, we ask: what happens, in an age of artificial intelligence, when scientists go *looking* for identity? In attempting to answer this question, we find that concerns about the *epistemic* and *identity-oriented* impacts of AI may converge in peculiar ways. We use two case studies to explore the use of AI in research on disability, sexuality, gender, and race. We argue that the cultural authority given to AI, combined with its increasing use in science, creates the possibility of revitalising and re-entrenching notions of identity and difference as "true," "objective" and "real". While this "materialisation" of difference is not a new concern, certain uses of AI make it a more urgent and an increasingly ubiquitous risk. We argue that researchers examining the social risks of AI must attend to the possibility of such materialisation. What would it look like for researchers to attend to the risk of materialisation? We suggest that such attentiveness requires historical awareness, and we point to the interdisciplinary literature around genomics—a technology that raises similar issues—as a valuable source and touchstone for scholars seeking to critically understand the epistemic impact of AI on questions of human identity.

1. Introduction

What happens when scientists go *looking* for identity with algorithmic tools? How is AI deployed to contextualise 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 hypothesised genetic origins

Corresponding Author: Os Keyes. Email: okeyes@uw.edu

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 revitalising and re-entrenching notions of identity and difference as "true," "objective" and "real."

While this materialisation 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 materialisation 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 dramatised 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 historicizing 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 specialised training, even for computer scientists" (Burrell 2016; Ananny and Crawford 2018). For McQuillan, 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 ritualised 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 center 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 valorise 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 et al. 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, postcolonial 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 summarises such worries in hypothesising 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" (Cheney-Lippold 2011, 177).

1.3. *Contextualising 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 hypothesised consequences of AI are neither specific to AI, nor particularly new. The "black-boxing" of decisions in science is understood as a long-standing 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 artifacts 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" (Waidzunas 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 emphasised 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, p.4) As Karin Knorr Cetina has documented, scientific cultures, and their approaches to technologies or knowledge, vary widely, between fields and individual laboratories (Cetina 2009). 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 and Bucher, who both take 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 contextualised and historicised. We are, of course, hardly the first people to make such a claim. Although her work is largely focused on surveillance and racialised 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 contextualise 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 et al. 2015; Samuels 2014).

In sum, then, we aim to historicise and contextualise concerns about the impact of AI on scientific practice, about the effect of AI on notions of identity, 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 legitimisation 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 naturalise a particular view of it—one that, unsurprisingly, conforms with status quo assumptions.

This process of naturalisation 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 characterises 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 stabilising 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 demedicalisation (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

[1]In 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,

and colleagues' "Whole-genome deep-learning analysis identifies contribution of non-coding 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" (Engineering and New 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 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 naturalises 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, this is a case study in how history and context informs this 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 organising and structuring their work. As Jennifer Singh has documented, autism research includes vast assemblages of researchers, state actors and (largely parent-driven) advocacy organisations. 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

improved truths inexorably overtaking old falsehoods (Verhoeff 2013; Hollin 2014)

2011). With autism, in particular, there is a strong neuroscientific and genomic bent within research as well as advocacy organisations, 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 19th 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 & Kozinski'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 generalised, 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

[2]As 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*.

[3]The 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.

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 naturalising identity. Rather, it 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 & Kosinksi 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; **?**). 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 legitimisation of individual lives. At the macro scale, these hypotheses have been used to justify queerphobia, due to the implication that queer people are biologically abberant, and to reinforce rigid ideas of gender and sexuality (Waidzunas 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 internalised 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 & Kozinski'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 et al. 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 21st century what "genetics" was to the 20th, or anthropometrics to the 19th—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 legitimise (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 method-

ologically) 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 legitimising 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 contextualise and historicise 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.

Acknowledgement(s)

Anonymised for review

Funding

Anonymised for review

Notes on contributor(s)

Anonymised for review

References

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20(3): 973–989.

Bautista MA, Hernández-Vela A, Escalera S, Igual L, Pujol O, Moya J, Violant V and Anguera MT (2015) A gesture recognition system for detecting behavioral patterns of adhd. *IEEE transactions on cybernetics* 46(1): 136–147.

Beer DG (2017) The social power of algorithms. *Information, Communication and Society* : 1–13.

Bhagat RB (2006) Census and caste enumeration: British legacy and contemporary practice in india. *Genus* : 119–134.

Bivens R and Haimson OL (2016) Baking gender into social media design: How platforms shape categories for users and advertisers. *Social Media+ Society* 2(4): 2056305116672486.

Bowker GC (2014) Big data, big questions— the theory/data thing. *International Journal of Communication* 8: 5.

Braun L (2014) *Breathing race into the machine: The surprising career of the spirometer from plantation to genetics*. U of Minnesota Press.

Brookey RA (2002) *Reinventing the male homosexual: The rhetoric and power of the gay gene*. Indiana University Press.

Browne S (2015) *Dark matters: On the surveillance of blackness*. Duke University Press.

Bucher T (2016) Neither black nor box: Ways of knowing algorithms. In: *Innovative methods in media and communication research*. Springer, pp. 81–98.

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 2053951715622512.

Calude CS and Longo G (2017) The deluge of spurious correlations in big data. *Foundations of science* 22(3): 595–612.

Calvo RA and D'Mello S (2010) Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1(1): 18–37.

Campolo A and Crawford K (2020) Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society* 6: 1–19.

Carrasquilla J and Melko RG (2017) Machine learning phases of matter. *Nature Physics* 13(5): 431.

Cetina KK (2009) *Epistemic cultures: How the sciences make knowledge*. 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, Vertesi J, Lynch ME and Woolgar S (2014) *Representation in scientific practice revisited*. 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, Morland I and Sullivan N (2015) *Fuckology: critical essays on John Money's diagnostic concepts*. University of Chicago Press.

Dumit J (2004) *Picturing personhood: Brain scans and biomedical identity*. Princeton University Press.

Ecker C (2011) Autism biomarkers for more efficacious diagnosis. *Biomarkers in medicine* 5(2): 193–195.

Elish MC and Boyd D (2018) Situating methods in the magic of big data and ai. *Communication monographs* 85(1): 57–80.

Engineering G and New B (2019) Ai finds autism-causing mutations in "junk" dna. *Genetic Engineering and Biotechnology News* URL `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*. St. Martin's Press.

Eyal G (2010) *The autism matrix*. Polity.

Felt U, Igelsböck J, Schikowitz A and Völker T (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*. University of Washington Press.

Fu S, He H and Hou ZG (2014) Learning race from face: A survey. *IEEE transactions on pattern analysis and machine intelligence* 36(12): 2483–2509.

Galison P and Daston L (2007) Objectivity .

Geertz C (1973) *Local knowledge: Further essays in interpretive anthropology*. Basic books.

Gillespie T (2014) The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167: 167.

Hames-Garcia MR (2011) *Identity complex: Making the case for multiplicity*. U of Minnesota Press.

Hashemi J, Spina TV, Tepper M, Esler A, Morellas V, Papanikolopoulos N and Sapiro G (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)*. IEEE, pp. 1–7.

Haslam N and Levy SR (2006) Essentialist beliefs about homosexuality: Structure and implications for prejudice. *Personality and Social Psychology Bulletin* 32(4): 471–485.

Hayes J, McCabe R, Ford T and Russell G (2020) Drawing a line in the sand: affect and testimony in autism assessment teams in the uk. *Sociology of Health & Illness* .

Helbing D, Frey BS, Gigerenzer G, Hafen E, Hagner M, Hofstetter Y, Van Den Hoven J, Zicari RV and Zwitter A (2019) Will democracy survive big data and artificial intelligence? In: *Towards Digital Enlightenment*. Springer, pp. 73–98.

Hird MJ (2004) *Sex, gender, and science*. Springer.

Hollin G (2014) Constructing a social subject: Autism and human sociality in the 1980s. *History of the Human Sciences* 27(4): 98–115.

Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH and Wenderoth N (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. *Real Life* 2.

Kitchin R (2017) Thinking critically about and researching algorithms. *Information, Communication & Society* 20(1): 14–29.

Krafft P, Young M, Katell M, Huang K and Bugingo G (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*. Harvard university press.

Law J (2004) *After method: Mess in social science research*. Routledge.

Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, Brewer D, Christakis N, Contractor

N, Fowler J, Gutmann M et al. (2009) Computational social science. *Science* 323(5915): 721–723.

Lugones M (2016) The coloniality of gender. In: *The Palgrave handbook of gender and development*. Springer, pp. 13–33.

L'heureux A, Grolinger K, Elyamany HF and Capretz MA (2017) Machine learning with big data: Challenges and approaches. *IEEE Access* 5: 7776–7797.

Maclure J (2019) The new ai spring: a deflationary view. *AI & SOCIETY* : 1–4.

Morandini JS, Blaszczynski A, Ross MW, Costa DS and Dar-Nimrod I (2015) Essentialist beliefs, sexual identity uncertainty, internalized homonegativity and psychological wellbeing in gay men. *Journal of counseling psychology* 62(3): 413.

M'charek A, Schramm K and Skinner D (2014) Topologies of race: Doing territory, population and identity in europe. *Science, Technology, & Human Values* 39(4): 468–487.

Natale S and Ballatore A (2017) Imagining the thinking machine: Technological myths and the rise of artificial intelligence. *Convergence* : 1354856517715164.

Noble SU (2018) *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

Ortega F and Choudhury S (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 Sengers P (2020) Making data science systems work. *Big Data & Society* 7(2): 2053951720939605.

Roberts D (2011) *Fatal invention: How science, politics, and big business re-create race in the twenty-first century*. New Press/ORIM.

Rose N and Abi-Rached JM (2013) *Neuro: The new brain sciences and the management of the mind*. Princeton University Press.

Samuels E (2014) *Fantasies of identification: Disability, gender, race*, volume 10. NYU Press.

Sato JR, Hoexter MQ, de Magalhães Oliveira Jr PP, Brammer MJ, Murphy D, Ecker C, Consortium MA et al. (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* URL `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 N (2019) Knowing algorithms. *Digital STS* : 412–422.

Singh JS (2015) *Multiple autisms: Spectrums of advocacy and genomic science*. U 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*. 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* URL `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*. Springer, pp. 176–183.

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*. U of Minnesota Press.

Waidzunas T and Epstein S (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 Kosinski M (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 CT and Paine D (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 KK, Wu Z, Bedbrook CN and Arnold FH (2018) Learned protein embeddings for machine learning. *Bioinformatics* 34(15): 2642–2648.

Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, Fak JJ, Funk J, Yao K, Tajima Y et al. (2019) Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nature genetics* 51(6): 973–980.

Zimmer F, Stock M, Scheibe K and Stock WG (2019) Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice* 7(2): 40–53.