Disabling AI: Biases and Values Embedded in Artificial Intelligence[1]

Dr. Damien Patrick Williams, Ph.D
University of North Carolina at Charlotte

Abstract: Algorithms and so-called artificial intelligence are embedded within society and human lives, and these fields’ directions hold major implications for both social and technological systems. I use multiple case studies to highlight how “AI” as it currently exists fails to account for the needs, experiences, and material conditions of multiple modes of human life. Then, drawing from the perspectives and research of women, disabled people, trans communities, Black people, Indigenous people, queer folx, and other marginalized identities, I describe an interdisciplinary program which better foregrounds the lived experiential knowledge of marginalized people. Finally, I argue that in order to redress the very real material harms of “AI” systems, we must ensure that the perspectives of marginalized communities are placed at the forefront and center of conversations about them, to help us radically rethink founding assumptions about what said systems are for.

Keywords: Algorithmic bias; Artificial intelligence; Disability studies; Ethics; Philosophy of technology; Race and gender studies

ORCID: 0000–0001–6652–2010

1. Introduction

The history of “artificial intelligence” consists largely of men who sought to gain access to an idealized pure and impersonal reason by removing emotion and affect from the idea of what it means to be a mind. From feminist epistemology, affective psychology, and computational scientific history came an increased understanding that removing “subjective” feeling was no way to create a whole mind, because emotions play a role in the very process of “knowing,” including within the minds and lives of the men trying to minimize that very role. As designers and activists embodying sites of marginalization based on disability, race, gender, sexuality, or combinations of the above gained prominence and built community, they highlighted the range of epistemic violence that is done by having to contort and shape oneself to fit into the demands and expectations of a disciplinary community. As a result, they have directly confronted

[1] This paper was excised and adapted from my 2022 Virginia Tech doctoral dissertation, *Belief, Values, Bias, and Agency: Development of and Entanglement with "Artificial Intelligence."* For a wider-ranging exploration of the variegated implications of “AI” history, please see that document and attendant citations.

assumptions about what being "right kind" of learner or knower means (Adam, 1998; Wilson, 2010; Dotson, 2011; Koopman, 2022).

Today, algorithmic "AI" tools exacerbate the pace and intensity of harms done by older, ingrained systems of oppression, whether in automated housing discrimination; facial recognition and predictive policing; the nonconsensual gathering and experimentation of medical and other data; or gene editing sold on the promise of "eliminating" disability and other "undesirable traits" (Hassein, 2017; Benjamin, 2019; Gilliard, 2019; The Takeaway, 2020; Brock, 2021). We find these harms in hiring systems excluding Black, disabled, female-identified candidates at higher rates; in grade management and online proctoring systems which work to entrain behavior without addressing students' core needs and challenges; in advertising algorithms capable of suggesting extremely specific products based on patterns gathered from online interactions; or in facial recognition systems which claim to be able to determine someone's sexual orientation (Gebhart, 2017; Biddle, 2018; Chen, et al., 2018; Quach, 2019; Brown, 2020; Grant-Chapman, Venzke and Quay-de la Vallee, 2020; Scherer and Shetty, 2021). Thus, we need to think in drastically different ways about how and why we build "algorithmic machine learning" applications, in the first place.

This chapter examines how the human beings who commission, design, build, and administer "AI" tools and systems come to embed their values, biases, and beliefs within what they create, and ends with recommendations for paths toward reducing the kinds of harm we do with and through these tools. In exploring various positions around and valences of power and lived experience which have driven the fields of "artificial intelligence" and "algorithmic machine learning," I deploy literature from philosophy of mind, philosophy of technology, cognitive science, "AI" engineering, and science and technology studies. This provides a firmer foundation from which to both highlight what kinds of tools and systems those conceptions have enabled—good and bad—and to then argue for intentionally disrupting and changing those tools and systems.

2. Matters of Perspective

In "ethical AI" literature and the public conception, "bias" is often used as though synonymous with "prejudice." However, a bias is simply a tendency or expectation towards a particular point of view or perspective. If one learns to match and anticipate patterns, then one is engaging in a perspectival bias. The problem, then, is not bias per se, but bias in service to uninterrogated prejudice or bigotry. Assumptions about gender, race, physical or mental ability are embedded in knowledge areas ranging from philosophy and data science to policing, employment practices, and even photography, and the actions taken by humans working under those assumption are

then translated into data which is used to train automated algorithmic systems.[2] In those instances, it is not the preference or expectation which does harm, but the unwillingness or inability to engage in "Bracketing"—a process of categorizing and accounting for the paths said expectations are likely to produce (Drew, 2004; Gearing, 2004; Charmaz, 2006; Starks and Trinidad, 2007; Tufford and Newman, 2016). Properly bracketed biases allow us to carefully consider our expectations and influences, providing us with the grounding by which to ask, "What questions haven't I asked?" or "Which perspectives have I failed to include?" These questions are crucial, because they help us to understand the dynamics of concepts within the disciplines of practice and training out of which "AI" algorithms are developed.

For reasons discussed elsewhere (Williams, D., 2012), I refrain from uncritically deploying the term "artificial intelligence" or "AI."[3] In a move which is distinct but conceptually connected to this discussion, the Georgetown Center for Privacy and Technology (CPT) announced they will no longer be using the terms "artificial intelligence," "AI," or "machine learning" on the basis that these terms obscure and often outright mislead the lay public as to the capabilities of the systems in place (Tucker, 2022). In recent years this obfuscation has been exacerbated by what is referred to as "criti-hype," a process whereby supposed critics of technology nonetheless uncritically reiterate many of the grandiose claims of the creators of these technologies.[4] However, where this 'term is usually used to suggest that anything with a certain amount of hype must then by default be negative, , I instead aim to critically and productively trouble both the intentional obfuscation performed by the terms "AI," "Machine Learning," etc., and the assumed notions of both "artificiality" and "intelligence." Subsequently, discussions or mentions of the term "Artificial Intelligence" and its variants are written in scare quotes.

3. Biases and Values Embedded in "AI"

To achieve the goal of creating ethically designed and administered "AI" systems which are beneficial to everyone, we must first ensure that they meet the needs of and redress the harms to the people most-often oppressed and marginalized. And in order to best incorporate these marginalized perspectives into "AI" research, we must understand what perspectives are *currently* embedded therein. Examining "AI" research today, we find it rife with harmful

[2] See also Code, 1991; Harding, 1991; Garry and Pearsall, 1996; Fredrickson et al., 1998;; Hobson, 2008; Abbate, 2012; Gürkan, 2015; English, et al., 2017; Hicks, 2017; Terrell, et al., 2017; McNeil, 2018; Cave and Dihal, 2020; Hanna, et al., 2020; and others explored in the next section.

[3] For different delineations of "naturalness," "artificiality," "intelligence," and "consciousness" through the Deweyan lens. cf. Flowers, J. (2019) Reconsidering the "Artificial," the "Intelligent," and the "Conscious" in Artificial Intelligence and Machine Consciousness through American Pragmatism. appearing in *Papers of the 2019 Towards Conscious AI Systems Symposium*.

[4] Cf. Lee Vinsel: https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5.

perspectives which are merely assumed to be true. Long-standing biases get encoded into new technology because the intended function of the new system is often modeled on the same or a similar set of assumptions already present in existing methods. Consequently, existing assumptions—and the values embodied ***by*** those assumptions—persist, and come to animate the new artifacts and systems of the newly "advanced" technologies. The perspectives and logics which get embedded in "AI" systems can be racist; they can be sexist; they can be predatorily capitalist; they can be ableist; and more. These embedded prejudices represent ways of thinking which have harmed real people, and have created and/or sustained systematic societal harms. And, as we are increasingly made aware, these systems' effects are disproportionately felt by already marginalized communities. The following overview is representative and will not be exhaustive.

4. Racist Values Embedded in "AI"

Racial prejudices have been found embedded in "AI" and algorithmic systems, including carceral systems which use surveillance techniques, facial recognition, and predictive metrics to determine policing tactics, sentencing, and punishments. In 2018, a demonstration of Amazon's Rekognition facial imaging system made false identification matches to 28 different congresspersons, linking their images to pictures found in mugshot databases (Snow, 2018). This outcome is entirely explicable. First, digital cameras and imaging systems see Black and brown faces—and particularly the faces of Black and brown women—less well (Buolamwini and Gebru, 2018). Next, most existing facial recognition systems are trained on mugshot databases. Third, as Black people are over-policed, with many young Black men being spuriously said to have "Fit the Description," the mugshots of Black people are entered into police databases at a disproportionate rate. Then, those databases are used to train algorithmic surveillance systems on how to search for "predictors of criminality." We have even taught these systems to search *for* particular skin tones (Joseph and Lipp, 2018).

Taken together, these algorithms thus apply pattern-recognition metrics in a manner typically resulting in darker skinned faces being marked, at a much higher rate, for suspected criminality. In effect, we have taught the algorithm how to ***automate*** the process of reinforcing the myth of Black people "fitting the description" (Garvie, et al., 2016; Burrington, 2018; Olson and Labuski, 2019; Williams, D., 2022). And these racialized prejudices are encoded not just in the surveillance state, but in the judgments made about people who are then subject to and made the subject of the justice system.

In a 2016 "Machine Bias" investigation, ProPublica demonstrated that the Compas algorithmic bail-setting and sentencing recommendation systems in use by Broward County, Florida were racially prejudicial (Angwin et al., 2016). The system might recommend that a Black man with no record of prior offenses and a lower likelihood of recidivism receive a lower likelihood of being granted bail and a harsher carceral sentence than a white man with priors and a higher likelihood of recidivism (Angwin et al., 2016). And this, again, is a result of the training data.

That is, historical data concerning human behavior that is decades, if not centuries old, is used to configure the parameters of these algorithmic "AI" systems, and in the process the "AI" then "learns" or replicates these patterns. (Mayeri, 2001; Brunson and Miller, 2006; Goff, et al., 2014; Ward, 2018; Hansen, 2019; Castle, 2021).

In his 2019 book *Black Software: The Internet and Racial Justice, from the AfroNet to Black Lives Matter*, Charlton McIlwain details (among other things) how IBM developed categorization systems used by Nazis during the holocaust as well as a comprehensive surveillance mechanism called the "Book Of Life" for Apartheid South Africa, deployed almost exclusively to track and monitor Black South Africans. The US government, seeing an opportunity to modernize the existing efforts of the Counterintelligence Program (COINTELPRO), asked IBM to develop automated tools to help them track, monitor, and respond to "the Black problem"—a euphemism for the direct and deliberate conflation of Black life with a perception of inherent criminality; thus the Law Enforcement Assistance Administration was formed (Churchill and Wall, 1990; McIlwain, 2019). In the 1960s and '70's COINTELPRO was tasked with monitoring and subverting Black Civil Rights leaders, from the Rev. Dr Martin Luther King, Jr. to the Black Panther Party. The high-level authorization and enactment of this surveillance program hinged upon the idea that these people—Black activists protesting for equal rights—were first and foremost worthy of being monitored, a belief stemming directly from the sociohistorical casting of Black people as inherently violent and "lesser."

The systems and techniques developed in this collaboration included weighted metrics of criminality to correlate the behavior and past records of anyone who had been included in these databases. At the time, this was done with paper records written by humans and encoded by hand into the system, and Black individuals and communities "somehow" always managed to be regarded as "higher risk incidents," thus rating increased response from police. Similarly, if a "high risk" (Black) individual was seen in a "low risk" (white) neighborhood, then that too would result in greater personnel deployment. These self-reinforcing risk categorizations would then be used as training data for the human police, and (eventually) their automated systems.

Even discounting avowed white supremacists, psychological researchers have demonstrated that Black children—and Black people in general—are almost always perceived as older and more imposing than white people of similar ages, heights, and builds, a fact which quite obviously has vast ramifications for Black peoples' encounters with the police. In the case of Black children, this can also result in white respondents having a hyper-sexualized perception of girls and an increased threat response toward boys. And this perception of Black people as more often violent, more often imposing, more often older than they actually are pervades not just the training procedures for how police are meant to respond to situations involving people of color, but also our popular culture, social media, and societal perceptions of Blackness, as a whole.

5. Gendered Values Embedded in "AI"

In their 2017 study, Caliskan, Bryson, and Narayanan demonstrated that Global Vectors for Word Representation (GloVe) and Word2Vec systems easily demonstrated correlations made along gendered lines between words like "King" and "Man," "Queen" and "Woman," "CEO" and "man," "secretary" and "woman," "doctor" and "man," "nurse" and "woman," "President" and "man," etc. (Caliskan, Bryson and Narayanan, 2017). Researchers also found pejorative and prejudicial associations between negative adjectives and "Black-sounding" names, thus marking the "Black-sounding" names as less pleasant and less employable than "white-sounding" ones (Bertrand and Mullainathan, 2004; Caliskan, Bryson and Narayanan, 2017). One reason—though by no means the only reason—for these associations is that the single most-often used cache of "machine learning" "natural language processing" training data is known as the "Enron Corpus," which is a cleaned and standardized collection of the over 600,000 emails between Enron executives which were entered into public record during discovery and prosecution of the Enron federal fraud case, in 2002. The Corpus' hundreds of thousands of emails contain millions of lines of natural language text between a very particular class and category of people, who talk in very specific, gendered, powered, and racialized ways about the topics under discussion.[5]

In addition to all of this, these systems are trained on publicly available word association studies, many of which demonstrate deep human prejudicial biases. These prejudices are not only present within vector-based word association systems, but also persist in pretrained transformers in an even more nuanced and systemic way (Williams, D. 2023a). Multiple user interactions with ChatGPT and Bard have demonstrated a persistent gendered prejudicial bias in associating the words "doctor," "lawyer," and "president" with the pronoun "he" and the words "nurse," "paralegal," and "secretary," with the pronoun "she." And those associations seem to be so heavily weighted that they are more likely to insist on grammatical errors before letting the hypothetical doctor, lawyer, or president have she/her pronouns. Run-throughs substituting the singular "they" returned responses purporting an inability to determine the subjects and objects in the sentence, until it was specified that the pronoun applied to the nurse, paralegal, or secretary— and then it "corrected" the pronoun to "she" in its responses. And yet, when asked to display the code or token weights it used to generate these outcomes, Bard declined, on the basis that doing so would generate something "discriminatory in nature,". Things only got worse when Google updated "Bard" to become "Gemini" (Gilliard, 2024).

6. Capitalist and Classist Values Embedded in "AI"

Benefits determinations systems are based on algorithms which use healthcare expenditure costs as proxies of healthcare outcomes, which can result in, e.g., poorer healthcare outcomes for marginalized populations (Obermeyer et al., 2019). Take the algorithms at work in the

[5] Cf. Google Scholar searches for "Enron Corpus" in 2021: https://web.archive.org/web/20210612231520/https://scholar.google.com/scholar?hl=en&as_sdt=0%2C47&q=%22enron+corpus%22.

Temporary Assistance for Needy Families (TANF) benefits systems, as showcased in Virginia Eubanks' 2018 book *Automating Inequality*. In this case, Eubanks demonstrates how people who are already at a lower socioeconomic status are made subject to systems that will keep them in poverty, rather than helping to elevate them out of poverty. And this is largely due to the kinds of assumptions that get embedded in the benefits system—assumptions about people's lives, about what kinds of needs they have, and about the "correct" purposes of the payouts they depend on.

Many have also proposed using "AI" systems for assessing risk levels for long-term houselessness (Denton, 2019), for matching the unhoused to available low-income homes (Khoo, 2019; Bishari, 2022), for paying parking tickets, or for finding available food to cut down on waste. But many of these "innovations" still tend to be created without the input, let alone direction of the communities they are intended to serve, meaning that they do not fully meet the needs of the people who may actually need them. These systems also require massive amounts of water and energy, increasing environmental costs which are and will likely continue to be unequally borne by marginalized communities around the world (Bender et al, 2021; O'Brien and Fingerhut, 2023). Unsurprisingly, then, we can also find ableist biases—and not just class-based ones—embedded in benefits systems, specifically those which make determinations about the kind of help and healthcare people need to live. These systems are not just opaque but have been trained on datasets which are, in many cases, filled with assumptions, including mistaken beliefs originating in nineteenth century notions about the forced institutionalization of disabled people.

7. Ableist Values Embedded in "AI"

At Georgetown Law School in January 2020, the "Strategic Advocacy on Disability Rights and 'AI' in Benefits Determinations" symposium was convened to discuss the use of automation and "AI" in disability benefits decisions across the United States. As highlighted in the symposium's ensuing policy report, algorithmic systems are being used in everything from assessment questionnaire systems, and electronic visit verification, to codifying able-bodied people's presuppositions about the "best environment" for a mythically monolithic category of disabled people (Brown et al., 2020). Many government-maintained lists of benefits distributions for the disabled and poor mandate certain styles of life and levels of income, meaning many recipients cannot get married to or even live with long-term partners, for fear of losing life-saving assistance (Social Security Administration, 2021).

The algorithmic frameworks used to determine how much of which kinds of assistance disabled people will have access to are all built from and trained on ableist notions of health and well-being which intersect with class, gender, and race (Obermeyer et al., 2019; Shew, 2020). Similarly, there is the use of robotics systems to "correct" autistic children's behaviour under the title of "Socially Assistive Robotics" or "Robot Augmented Therapy," processes which are almost universally undertaken in ways that egregiously dehumanize neurodivergent children and adults (Williams, R., 2021b). There are also multiple cases involving automated vehicles where

the vision systems fail to properly categorize wheelchair users or people using crutches as pedestrians (Hao, 2018).

Additionally, many fields either already have or are planning to integrate pretrained transformers into their public-facing interface, such as healthcare providers promising "AI"-enabled guidance, or the automated plagiarism checker Turnitin announcing new GPT-checking tools to try to "catch" students in new wave of "AI"-assisted cheating. But user explorations of pretrained transformer tools have yielded results which mirror the ableism present in many previous tools and devices. When prompted to help guide disability benefits determinations, Bard consistently offers more money for live-in facility care than in-home care, repeating previous decisions concerning benefits determinations. And as researchers such as Rua Williams and Janelle Shane first demonstrated and others have corroborated, the new crop of ChatGPT detectors being applied to student work have problems not only with text written by non-native English speakers (Liang et al., 2023), but also by neurodivergent individuals.[6] Integrating these capabilities into automated plagiarism and proctoring software, which already endangers disabled and otherwise marginalized students, will likely only exacerbate the problem. When considering the impact and consequences of automated ableism in this way, we may come to understand that ableism not just acute but often a pervasive background condition of everyone's lives.

8. Intersectional Oppression Embedded in "AI"

In her 2018 book *Algorithms of Oppression*, Safiya Noble discusses how stereotypical perceptions of Black people, especially women and girls, are rendered in American culture through the lenses of Google's search and advertising metrics, returning search results that often reflect and reinforce those same stereotypes (Noble, 2018). To understand this, we must first remember that, in addition to the paid advertisements described by Noble, Google delivers search results through a three-part process they describe as "Crawling, Indexing, and Serving/Ranking."[7] First, automated Google systems continually trawl the internet for the most current versions of webpages. Second, those processes work to cross-reference the content of that page and get a sense of what it is. Finally, Google uses factors such as the searcher's location, language, and type of device to decide the order by which it will serve up the results it has generated.

None of this is a "neutral" or "objective" process. Every step depends on and changes via choices made by both the searcher and the developers who designed the search system. Every one of those factors—e.g. location, device type, language choice, and the weight they are given—impacts what the end user receives. Your location, for instance, will be used to give heavier weight to whichever links other users near you have selected when they have done

[6] Cf. Janelle Shane "Apparently I Am A Robot" https://www.aiweirdness.com/writing-like-a-robot/; Rua Williams, "oh no I am a robot" https://kolektiva.social/@FractalEcho/109480110824287470.

[7] "How Google Search Works" https://support.google.com/webmasters/answer/70897?hl=en.

similar searches. And if one leaves Google's autocomplete feature enabled, all of this begins before you finish typing your query, as the algorithm presents you with a list of pre-determined options along with previews of the attendant results. Thus, if web content providers have paid to have their content served as ads and those ads are more often clicked on by users in your location, then those results will be served to you not just as ads but as supposedly "neutral" search results.[8]

Several corporations are developing health-related technologies, combining many of the "AI" capabilities mentioned thus far, e.g. brain-computer interfaces (BCIs) such as Elon Musk's proposed NeuraLink and the Wear OS biometric system from Google and Samsung. Both innovations are meant to be full-suite biometric readers which can monitor your brain states, your autonomic system, your blood-oxygen levels, and your level of hydration, all in real-time. But these systems are also meant to do things like monitor your gait to warn you if you are about to fall, track your resting heart rate and perspiration, and monitor your vocal tone in conversations, to recommend whether you might want to modulate your tone to be better perceived by your interlocutors. Unfortunately, gait monitors that have been trained on non-disabled ambulatory users will not accurately recognize the gait of someone with spina bifida or cerebral palsy. And voice recognition software historically has difficulty with speech patterns of disabled users whose disability affects their speech. And this is before we consider the ways health monitoring apps, BCIs, and other "AI" tools could be leveraged to track and control populations in schools, the workplace, or even in the privacy of one's home.[9]

In Western culture, and the United States in particular, Black people are more harshly scrutinized and judged as regards their emotional comportment in social situations. And Black women's vocal tones, in particular, are often policed for how they interact with each other and present themselves in conversation. The result is that Black women are often told that they're being overly agitated or angry, even when they are presenting neutrally, or more harshly judged for actually being angry, even when they have every right to be (Smith and Moore, 2019; Owens 2020). For a system like Wear OS to make judgements about vocal tone, it will have to be trained on conversational inputs. If the training data is limited to the kinds of everyday interactions that white western male human programmers and designers assume is "correct" and "proper," then these assumptions will likely be embedded in these tracking systems.

Emotive content in vocal tone is inherently cultural, and if the people who design and program the tools do not account for the kind of inherent biases towards certain types of comportment,

[8] If you want to change this, go to Google's "Settings" to adjust options for search language, search history, and what search data of yours Google admits retaining. The pertinent category is "Search Settings," then and "Private Results," "Search History," and "Region Settings." Again: altering any of these will alter what Google shows you and how.

[9] Cf. Brown et al., 2022

expression, lived experience, and behavior, those expectations will be replicated in these biometric tools (Shetty and Quay-de la Vallee, 2022). If these systems, for example, suggest to Black users that, “you might want to calm down,” they run the risk of instantiating the stereotype of the “angry black man” or “angry black woman” into the device and its operations. And if the health assessment applications of these systems are also primarily trained on healthcare interactions from white users, then they have a high likelihood of simply reproducing and perpetuating the erroneous assumptions and prejudices about “health” that Black people already have experienced in human interactions (Hoffman et al., 2016). Disabled and neurodivergent Black people, then, will be particularly ill-served by these algorithmic solutions. And these intersectional prejudices have also been documented in computer vision and algorithmic facial recognition systems, where ableist, sexist, transphobic, and racist vectors are a component of the intersectional problems encoded in the system (Collins, 2012; Hoffmann, 2017; Buolamwini and Gebru, 2018; Scheuerman, Paul and Brubaker, 2019).

But racism is not limited to Western cultures, and histories of colorism, colonialism, and class stratification have led different formulations of anti-Blackness being found (alongside more local forms of racism) in China, Japan, and Korea—current hubs of “AI” and algorithmic research (Kim, 2015; Ouassini et al., 2021; Kanesaka, 2022). Several researchers have recently claimed to be able to use facial recognition and biometrics to determine and predict everything from homosexuality to criminal behaviour, ignoring the fact that in some parts of the world these are perceived to be the same thing and are punishable by death (Wu and Zhang 2016; Hao 2018; Farivar 2018; Quach 2019). “AI” health researchers recently prompted their system to find molecular compounds more efficient at harming humans, and were shocked when it did just exactly that, even going so far as to comment that “the thought had never previously struck [them]” that this might be possible and that they “are not trained to consider it” (Urbina, et al., 2022). Similarly, many have long worried about lethal autonomous weapons systems (LAWS)—or what have also been called “killer drones”—and their capabilities to autonomously differentiate military from civilian targets (Robbins, 2016). And yet governments and militaries give similar systems names like “the Gospel” in a play to bless human-trained decision-making systems with divine infallibility (Williams, D. 2023b; Suchman, 2024).

What we are seeing is not the product of a rogue machine going off the leash. These are standard operating procedures. In effect, these racist, sexist, and otherwise prejudiced assumptions have been encoded into a whole spectrum of hi- and low-tech developments, from tools as seemingly simple as pinhole cameras all the way to systems as complex as modern-day “artificial intelligent” résumé sorters. When asked to perform a task, “AI” and “machine learning” often provide outputs which seem surprising to humans, but they are in fact doing exactly what the systems were built, trained, and asked to do. They discover statistically validated patterns in the data on which they were trained, adjust the weighted connections in their neural networks, and return outputs which correspond to the provided inputs. The systems are simply following parameters that we gave them.

9. Conclusion: The Role of Marginalized Perspectives in "AI" Ethics

Here, I advocate a new understanding of "AI" ethics which starts with the technology's intersection with biosocial systems of knowledge, power, values, and beliefs and starting from the vantage point of those people who have been doing this work the longest. We must interrogate the constructions of identity which presuppose a "right kind" of knower or learning subject, or type of person whose experiences can even count as legitimate. Certain performances of whiteness and ablebodiedness cut across beliefs, political power, knowledge formation, and medicalization, and how combinations of these make their way into technological systems and artifacts (Williams, D., 2022). All technosocial tools and systems, including and especially "AI", replicate, instantiate, and iterate upon the assumptions and the values of the people who have commissioned them, programmed them, and trained them.

The harmful outcomes of algorithmic systems and tools arise from the replication, reinforcement, iteration, and exacerbation of already extant racist, sexist, ableist, transphobic, homophobic, fatphobic, and otherwise bigoted human values. And the ways these systems interact when they operate in the world are derived from, and generate anew, the data on which they train. Thus, we must work to recognize, unpack, and challenge assumptions about the world; seek diverse and even unfamiliar ways of knowing. This will require us to research, devise, and teach new models for the creation of knowledge and ethical engagement; to learn how to agilely deploy these models, placing them in conversation with each other; and build intersubjective bases for shared understanding. All of which means there is no one-size-fits-all answer, only a shifting matrix of needs, stakeholders, rightsholders, and power dynamics.

Though academic and industry calls for "Algorithmic Transparency" and "ethical alignment" of "AI" systems are well-intentioned, many presuppose either that the general populace is merely coincidentally uninformed about these systems, or that the creators and designers innocently forget to consider socioethical implications. Neither of these assumptions get at the crux of the issue, which is that corporations are both legally and morally responsible for the harms their "AI" cause, via intent or neglect. And these corporations continue to build "AI" even when, as we have now established, the harmful outcomes of their operations were clearly foreseeable. Not only this, but the means to understand these systems are often intentionally ciphered and hidden from us, and thus the ethical implications actively obscured or ignored.

An example: When Bard was asked to provide code which would run the heavily prejudicial word associations it gave, above, it balked. It gave me something like that code, but it would not describe its own previously generated weights as part of the "associated"/"not associated" determination functions, because the resulting output would have fit its definition of a "discriminatory output." In order to get it to provide the lines of code with the "associated"/"not associated" indicated, I had to change the parameters to "innocuous" words such as "canid," "feline," "feral," "zoo," "housepet," and "animorphs," and then plug the weighted distribution for the corresponding "discriminatory" words in their place. This worked just fine. When public-

facing LLM's are adjusted to filter out prejudicial biases *post hoc*, those "fixes" amount to little more than band-aids at best. We know and have demonstrated that the system more heavily privileges certain stereotypical relationships, but "fixing" that fact by preventing the system from show us how those outcomes were derived is not, in fact, a fix. Instead, it is the equivalent of building an "AI" system which embodies the belief that "talking about discrimination is the real discrimination."

Note, however, that none of this is dispositive. In fact, it cannot be. That is because neither I nor any other end-user has access to the system's training data and the resulting weighted connections; all we have are the inputs we provide, and the outputs produced by the algorithm. This is why such "AI" systems are often called "black boxes." Aside from raising the alarm, there is little that can be done about it. To this end, knowing how these black box systems work—i.e., learning how they operate and what they do—is certainly important, but that knowledge means nothing if we cannot identify the source of the problem and meaningfully enforce changes in the design, construction, and implementation of these products. This is especially true as big-tech companies repeatedly move to eliminate what little internal oversight they have, and limit access independent researchers have managed to scrape together (Angwin et al., 2016; Vallor, 2016; Pasquale, 2017; Faife, 2021; Głowacka and Iwańska, 2021). If technologically solutionist projects are undertaken without first examining the oppressive logics at their root, then all that is likely to change is who commits the oppression.

Increasingly, people working within "AI Safety" call for moratoria or "pauses" on the development of LLMs and pretrained transformers that are "more powerful" than the current offerings from OpenAI and Google. Even Geoffrey Hinton, one of the pioneers of today's "AI," quit his position at Google so that he could speak out against the very tools he helped create. These calls to slow down or even halt development of these tools are rooted in the idea that said tools may one day become "smarter" than humans, and then destroy us all—but, it should be noted, none of them have asked why such a thing would even be possible, let alone pursued by any suddenly conscious "AI." As demonstrated above, these systems are often constructed with and trained on logics of carcerality, militarism, extractive capitalism, and white supremacy. They, therefore, reinforce extant assumptions regarding privilege, power, and profit. And even if we were to concede the point that we are on a path toward "human-like" "AI" or even "Artificial General Intelligence" (AGI), then the above concerns would matter even more. Because, as argued elsewhere, why would you want to create a mind from these values and with these morals (Gunkel, 2012; Estrada, 2019; Williams, D. 2019)?

But even as these facts stand in ever starker relief, the most obvious potential remedy is still all-too-often overlooked. Namely, centering the experiences of those individuals and communities who have been marginalized by existing systems of oppression. In doing so, we can not only better recognize how marginalized perspectives have been excluded from technology, but also begin to develop the tools, systems, and educational processes for reimagining research into algorithms, "AI," and other technologies. This approach provides us with a model for further

integration of marginalized and minoritized populations—what Gayatri Spivak calls "the Subaltern"—and providing them with primacy of place within organizations, as well as meaningful oversight and authority on those issues which most often affect them and their communities (Spivak, 1988; Claypool, et al., 2021). However, it should not be expected that those who have been most harmed by these systems devise the means to redress their harms. That expectation is (in itself) another oppressive harm.

Making the needed changes will mean integrating perspectives from fields like disability studies, philosophy, sociology, and science and technology studies—not only providing opportunities for the "sublaltern to speak," but learning how to listen to and take responsibility for what comes to be said. We must, therefore, center the experiential knowledge of disabled and neurodivergent "AI" researchers who unflinchingly call into question the assumptions about "standard" minds and bodies which have made their way into "AI" research over the past century, and suggest crucial remedies (Zebrowski, 2017a, 2017b, 2017c; Williams, R., 2018, 2021a, 2022; Ymous et al, 2020, Sum et al., 2022). Similarly, interventions from disabled designers and activists who reveal and examine how their particular experiences have been excluded from technology, and who can, instead, develop new research into and relationships with algorithms and other technologies (Wong, 2013; Kane, 2016; Jackson and Haagaard, 2020;).

This is not, we must stress, an "ethics-washing" afterthought once something goes wrong, or even hiring a team of internal ethicists to serve on seemingly powerless "Ethics Boards" who can then be dismissed when the insights they provide prove embarrassing or unprofitable (Wagner, 2018; Simonite, 2020; BBC, 2021; Bender et al., 2021). If we desire to change how "AI" systems are instantiated, entangled, and perpetuated in our cultures and to stop the variously bigoted and oppressive outcomes they have so far produced, then we need meaningful change to the training datasets, the development and design teams, the managerial principles of the corporation, the education and leadership of CEOs, the funding sources, the research questions, and the aims, beliefs, and values of which they're made.

Fortunately, there is some evidence that the adjustments recommended here are beginning to be made, both domestically and internationally.[10] But too often these recommendations fall short of identifying the systemic nature of prejudicial, oppressive, or otherwise harmful social elements, and instead limit discussion of "bias" to matters of personal responsibility. But we can change this, by ensuring that the perspectives and lived experiences of marginalized people are heeded in conversations about the design and implementation of algorithmic applications, even and perhaps especially when those perspectives make us uncomfortable. We must continually ask, who is in the room when we make the decisions that influence, shape, or even determine research

[10] Cf. Venable et al., 2016; Hamraie and Fritsch, 2019; Appleton, 2019; Whittaker et al., 2019; Lewis, et al., 2020; Milner and Traub, 2021; United Nations, 2021a; Kaye, 2022a, 2022b; State Internet Information Office, et al., 2022; Xióng Jié, 2022;Biden, 2023; the EU, and others.

directions? Who is missing from those rooms? Who finds themselves “the only one in [those] rooms?” Who determines what questions get asked, and shapes the kinds of answers obtained? Putting disabled and otherwise marginalized people with lived experiential expertise at the forefront of our conversations about “AI” may require us to radically rethink our founding assumptions about what “AI” and automation are for. But for millions of people, doing so could mean the difference between life and death.

References

Abbate, J. (2012). *Recoding Gender: Women's Changing Participation in Computing*. MIT Press.

Adam, A. (1998). *Artificial Knowing: Gender and the Thinking Machine*. Routledge.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Appleton, N. S. (2019). Do Not "Decolonize"… If You Are Not Decolonizing: Progressive Language and Planning Beyond a Hollow Academic Rebranding. *Critical Ethnic Studies*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FaccT '21). Association for Computing Machinery.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for The New Jim Code*. Polity.

Bertrand, M., Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha And Jamal? A Field Experiment On Labor Market Discrimination. *American Economic Review*, v94(4 Sep), 991–1013.

Biddle, S. (2018). Facebook Allowed Advertisers to Target Users Interested in "White Genocide"—Even in Wake of Pittsburgh Massacre. *The Intercept*.

Biden, J. R. (2023). Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. Federal Register. https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.

Bishari, N. (2022). San Francisco Rations Housing by Scoring Homeless People's Trauma. By Design, Most Fail to Qualify. *The San Francisco Public Press*.

Brock, A. L. (2021). On Race and Technoculture. *Microsoft's Race and Technology Research Lecture Series*.

Brown, L. (2020). How Automated Test Proctoring Software Discriminates Against Disabled Students. *Center for Democracy and Technology*.

Brown, L., Richardson, M., Shetty, R., Crawford, A. (2020). Report: Challenging the Use of Algorithm-driven Decision-making in Benefits Determinations Affecting People with Disabilities. *Center For Democracy and Technology*.

Brown, L., Shetty, R., Scherer, M., Crawford, A. (2022). Ableism And Disability Discrimination In New Surveillance Technologies: How new surveillance technologies in education, policing,

health care, and the workplace disproportionately harm disabled people. *Center For Democracy and Technology*.

Browne, S. (2015) *Dark Matters*. Durham, NC: Duke University Press, 2015.

Brunson, R. K., and Miller, J. (2006). Young black men and urban policing in the United States. *British Journal of Criminology*. 46.4.

Buolamwini, J., Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of *Machine Learning Research* 81.

Burrington, I. (2018). Policing Is an Information Business. *Urban Omnibus*.

Caliskan, A., Bryson, J. J., Narayanan, A. (2017). Semantics Derived Automatically From Language Corpora Contain Human-Like Biases. *Science.*

Castle, T. (2021). “Cops and the Klan”: Police Disavowal of Risk and Minimization of Threat from the Far-Right. *Critical Criminology* 29, no. 2.

Cave, S., Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology* 33.4.

Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE Publications.

Chen, L., Ma, R., Hannák, A., Wilson, C. (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Churchill, W., Wall, J. V. (1990). *The COINTELPRO Papers*. South End.

Claypool, H., Carey, C., Hart, A. C., Lassiter, L. (2021). Centering Disability in Technology Policy: Issue Landscape and Potential Opportunities for Action. *Center for Democracy and Technology*; *American Association of People with Disabilities*.

Code, L. (1991). *What Can She Know? Feminist Theory and the Construction of Knowledge*. Cornell University Press.

Collins, P. H. (2012). Social inequality, power, and politics: Intersectionality and American pragmatism in dialogue. *The Journal of Speculative Philosophy* 26.2.

Denton, J. (2019). Will Algorithmic Tools Help or Harm the Homeless? *Pacific Standard*.

Dotson, K. (2011) Tracking epistemic violence, tracking practices of silencing. *Hypatia* 26, no. 2.

Drew, N. (2004). Creating a Synthesis of Intentionality: The Role of the Bracketing Facilitator. *Advances in Nursing Science* 27(3).

English, D., Bowleg, L., Del Río-González, A. M., Tschann, J. M., Agans, R. P.; Malebranche, D. J. (2017). Measuring Black men’s police-based discrimination experiences: Development and validation of the Police and Law Enforcement (PLE) Scale. *Cultural Diversity & Ethnic Minority Psychology*. Vol. 23,2.

Estrada, D. (2020). Human supremacy as posthuman risk. *Journal of Sociotechnical Critique*, *1*(1).

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin’s Press.

Faife, C. (2021). Facebook Rolls Out News Feed Change That Blocks Watchdogs from Gathering Data. *The Markup*.

Farivar, C. (2018). “Central Londoners to be subjected to facial recognition test this week.” *Ars Technica*.

Fredrickson, B. L., Roberts, T.-A., Noll, S. M., Quinn, D. M., Twenge, J. M. (1998). That swimsuit becomes you: Sex differences in self-objectification, restrained eating, and math performance. *Journal of Personality and Social Psychology*, *75*(1).

Garry, A., Pearsall, M. (1996). *Women, Knowledge, and Reality: Explorations in Feminist Philosophy*. 2nd ed. Routledge.

Garvie, C., Bedoya, A., Frankle, J. (2016). The Perpetual Line-up: Unregulated Police Face Recognition in America. *Georgetown Law’s Center on Privacy & Technology*.

Gearing, R. (2004). Bracketing in Research: A Typology. *Qualitative Health Research* 14(10).

Gebhart, G. (2017). Privacy By Practice, Not Just By Policy: A System Administrator Advocating for Student Privacy. *Electronic Frontier Foundation.*

Gilliard, C. (2019). Banking on Your Data: the Role of Big Data in Financial Services. Prepared Testimony And Statement For The Record before the House Financial Services Committee Task Force on Financial Technology.

Gilliard, C. (2024). The Deeper Problem With Google’s Racially Diverse Nazis. *The Atlantic*.

Głowacka, D., Iwańska, K. Algorithms of trauma: new case study shows that Facebook doesn’t give users real control over disturbing surveillance ads. *Panoptykon*.

Goff, P. A., Jackson, M. C., Allison, B., Di Leone, L., Culotta, C. M., DiTomasso, N.A. (2014). The Essence of Innocence: Consequences of Dehumanizing Black Children. *Journal of Personality and Social Psychology*.

Grant-Chapman, H., Venzke, C., Quay-de la Vallee, H. (2020). A Year in Review: Student Privacy Issues Through a Season of Unprecedented Change. *Center for Democracy and Technology*.

Gunkel, D. (2012). *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. The MIT Press.

Gürkan, E. B. (2015). Women's Bodies as First Colony: A Study in the Hybrid Feminist Personal.

Hamraie, A., Fritsch, K. (2019). Crip technoscience manifesto. *Catalyst: Feminism, Theory, Technoscience*, *5*(1).

Hansen, Chelsea. (2019). Slave Patrols: An Early Form of American Policing.

Hanna, A., Denton, E., Amironesei, R., Smart, A., Nicole, H. (2020). Lines of Sight. *Logic Magazine*.

Hao, K. (2018). Can You Make an AI That Isn't Ableist? *MIT Technology Review*.

Harding, S. (1991). *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Hassein, N. (2017). Against Black Inclusion in Facial Recognition. *Digital Talking Drum*.

Hicks, M. (2017). *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge In Computing*. MIT Press

Hobson, J. (2008). Digital whiteness, primitive blackness: Racializing the "digital divide" in film and new media. *Feminist Media Studies*, *8*(2).

Hoffman, K. M., Trawalter, S., Axt, J. R., Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, *113*(16).

Hoffmann, A. L. (2017). Data, technology, and gender: Thinking about (and from) trans lives. In *Spaces for the Future*. Shew, A., Pitt, J. C., eds. Routledge.

Jackson, L., Haagaard, A. (2020). Designers in Residence. *Design for Social Innovation*. https://vimeo.com/470787270.

Joseph, G., Lipp, K. (2018). IBM used NYPD surveillance footage to develop technology that lets police search by skin color. *The Intercept*.

Kane, N. D. (2016). ‘Means Well’ Technology Technology and the Internet of Good Intentions. *Medium*.

Kanesaka, E. (2022). Racist Attachments: Dakko-chan, Black Kitsch, and Kawaii Culture. *Positions: asia critique*, *30*(1).

Kaye, K. (2022a). The FTC’s new enforcement weapon spells death for algorithms. *Protocol*.

Kaye, K. (2022b). How to kill an algorithm. *Protocol*.

Khoo, C. (2021). Comments of Georgetown Center on Privacy & Technology in response to Draft NIST Special Publication 1270: A Proposal for Identifying and Managing Bias in Artificial Intelligence. *The National Institute of Standards and Technology*.

Kim, S. J. (2015). Korean-origin kindergarten children’s response to African-American characters in race-themed picture books. *Education Research International*, *2015*.

Koopman, K. A. (2022). *Who Speaks Truth to Fiction? Scientific Authority and Social Difference in Speculative Fiction* (Doctoral dissertation, Virginia Tech).

Lewis, J. E., Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., Duncan, K., Garzon, S., Harrell, D. F., Jones, P-L., Kealiikanakaoleohaililani, K., Kelleher, M., Kite, S., Lagon, O., Leigh, J., Levesque, M., Mahelona, K., Moses, C., Nahuewai, I. (‘Ika’aka), Noe, K., Olson, D., Parker Jones, ‘Ōiwi, Running Wolf, C., Running Wolf, M., Silva, M., Fragnito, S., and Whaanga, H. (2020). *Indigenous Protocol and Artificial Intelligence Position Paper*.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers.

Margaret Mitchell: Google fires AI ethics founder. (2021). *BBC*.

Mayeri, S. (2000). A common fate of discrimination: Race-gender analogies in legal and historical perspective. *Yale LJ*, *110*, 1045.

McIlwain, C. D. (2019). *Black software: The Internet and racial justice, from the AfroNet to Black Lives Matter*. Oxford University Press, USA.

McNeil, Joanne. (2018). Big Brother’s Blind Spot. *The Baffler*, (40).

Milner, Y., Traub, A. (2021). Data capitalism and algorithmic racism.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

O’Brien, M., Fingerhut, H. (2023). A.I. Tools Fueled A 34% Spike in Microsoft’s Water Consumption, and One City with Its Data Centers Is Concerned About the Effect on Residential Supply. *Fortune*.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464).

Olson, P., Labuski, C. (2018). ‘There’s always a [white] man in the loop’: The gendered and racialized politics of civilian drones. *Social Studies of Science*, *48*(4).

Ouassini, A., Amini, M., Ouassini, N. (2022). # ChinaMustexplain: Global Tweets, COVID-19, and Anti-Black Racism in China. *The Review of Black Political Economy*, *49*(1).

The Overlooked Reality of Police Violence Against Disabled Black Americans. (2020). *The Takeaway*.

Owens, M. A. (2020). Closet Chair and Committee Side Piece. *Presumed Incompetent II: Race, Class, Power, and Resistance of Women in Academia*. Niemann, Yolanda Flores, Gutiérrez y Muhs Gabriella, and Carmen G Gonzalez, eds.

Pasquale, F. (2017). Toward a fourth law of robotics: Preserving attribution, responsibility, and explainability in an algorithmic society. *Ohio St. LJ*, *78*.

Provisions on the Administration of Algorithm Recommendations for Internet Information Services. (2022). State Internet Information Office; Ministry of Industry and Information Technology of the People’s Republic of China; Ministry of Public Security of the People’s Republic of China; State Administration for Market Regulation. (Original in Chinese, translated via Google.)

Quach, K. (2019). The infamous AI gaydar study was repeated–and, no, code can’t tell if you’re straight or not just from your face. *The Register*.

Scherer, M., and Shetty, R. (2021). NY City Council Rams Through Once-Promising but Deeply Flawed Bill on AI Hiring Tools. *Center for Democracy and Technology*.

Scheuerman, M. K., Paul, J. M., Brubaker, J. R. (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW).

Shew, A. (2020). Ableism, 20echnoableism, and future AI. *IEEE Technology and Society Magazine*, *39*(1).

Shetty, R., Quay-de la Vallee, H. (2022). CDT Comments to OSTP Highlight How Biometrics Impact Disabled People. *Center for Democracy and Technology*.

Simonite, T. (2020). A Prominent AI Ethics Researcher Says Google Fired Her. *WIRED*.

Sins Invalid. (2015). 10 Principles of Disability Justice. *Sins Invalid*.

Smith, M. D.; Moore, M. N. (2019). Black Feminist Thought: A Response to White Fragility. In *Black Women and Social Justice Education: Legacies and Lessons*, Stephanie Y. Evans, et al., eds. State University of New York Press.

Snow, J. (2018). Amazon’s Face Recognition Falsely Matched 28 Members of Congress With Mugshots. *Technology & Civil Liberties Attorney, ACLU of Northern California*.

Social Security Administration (2021) Understanding Supplemental Security Income SSI Eligibility Requirements. The United States Social Security Administration.

Solon, O. (2019). Why did Microsoft fund an Israeli firm that surveils West Bank Palestinians? *NBC News*.

Spivak, G. C. (1988). Can the Subaltern Speak? *Die Philosophin*, *14*(27).

Starks, H. and Trinidad, S. B. (2007). Choose Your Method: A Comparison of Phenomenology, Discourse Analysis, and Grounded Theory. *Qualitative Health Research* 17(10).

Suchman, L. (2024). The Algorithmically Accelerated Killing Machine. *AI Now Institute*.

Sum, C. M., Alharbi, R., Spektor, F., Bennett, C. L., Harrington, C., Spiel, K., and Williams, R. M. (2022). Dreaming Disability Justice in HCI.

Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., Parnin, C., Stallings, J. (2017). Gender differences and bias in open source: pull request acceptance of women versus men. *PeerJ Computer Science*. 3:e111

Tucker, E. (2022). Artifice and Intelligence. *Center on Privacy & Technology at Georgetown Law Blog*.

Tufford, L., and Newman, P. (2012). Bracketing in Qualitative Research. *Qualitative Social Work* 11, no. 1.

United Nations. (2021). Artificial intelligence risks to privacy demand urgent action – Bachelet. *Office of the High Commissioner for Human Rights*.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence* 4, no. 3.

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a World Worth Wanting*. Oxford University Press.

Venable, J., Sato, B. A., Del Duca, J., and Sage, F. (2016). Decolonizing Our Own Stories: A Project of the Student Storytellers Indigenizing the Academy (Ssita) Group. *International Review of Qualitative Research* 9, no. 3.

Wagner, B. (2018). Ethics as an Escape From Regulation. From “Ethics-Washing” To Ethics-Shopping? In *BEING PROFILED: COGITAS ERGO SUM: COGITAS ERGO SUM: 10 Years of Profiling the European Citizen*, Bayamlioglu, E, Baraliuc, I., Janssens, L. A. W. and Hildebrandt, M. eds, Amsterdam University Press.

Ward, G. (2018). LIVING HISTORIES OF WHITE SUPREMACIST POLICING: Towards Transformative Justice. *Du Bois Review: Social Science Research on Race* 15, no. 1.

Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., West, S. M. (2019). Disability, Bias, and AI. *AI Now Institute*.

Williams, D. P. (2012). Strange Things Happen at the One Two Point: The Implications of Autonomous Created Intelligence in Speculative Fiction Media. *THE MACHINE QUESTION: AI, ETHICS AND MORAL RESPONSIBILITY.*

Williams, D. P. (2019a). Consciousness and Conscious Machines: What’s At Stake?. In *AAAI Spring Symposium: Towards Conscious AI Systems.*

Williams, D. P. (2022). *Belief, Values, Bias, and Agency: Development of and Entanglement with Artificial Intelligence* (Doctoral dissertation, Virginia Tech).

Williams, D. P. (2023a). Bias Optimizers. *American Scientist*, 111(4).

Williams, D. P. (2023b). “Any Sufficiently Transparent Magic…” *American Religion*, 5(1)

Williams, R. M. (2018). “Autonomously Autistic: exposing the locus of autistic pathology,” *Canadian Journal of Disability Studies*, vol. 7, no. 2.

Williams, R. M. (2021a). Six Ways of Looking at Fractal Mechanics. *Catalyst: Feminism, Theory, Technoscience*, *7*(2).

Williams, R. M. (2021b). I, misfit: Empty fortresses, social robots, and peculiar relations in autism research. *Techné: Research in Philosophy and Technology*, *25*(3).

Williams, R. M. (2022). All Robots Are Disabled. In *Social Robots in Social Institutions*. IOS Press

Wilson, E. (2010) *Affect and Artificial Intelligence.* Seattle: University of Washington Press.

Wong, A. The Disability Visibility Project. 2013—Present. https://disabilityvisibilityproject.com/about/.

Wu, X., Zhang, X. (2016). Automated Inference on Criminality using Face Images. arXiv:1611.04135

Xióng Jié. (2022). Algorithmic Recommendations Can Finally Be Turned Off, China’s Provisions are a World First. *Guancha.cn*. https://interpret.csis.org/translations/algorithmic-recommendations-can-finally-be-turned-off-chinas-provisions-are-a-world-first/.

Ymous, A., Spiel, K., Keyes, O., Williams, R. M., Good, J., Hornecker, E., Bennett, C. L. (2020, April). "I am just terrified of my future"—Epistemic Violence in Disability Related Technology Research. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* .

Zebrowski, R. (2017a). An Android’s Dream: On Bodies, Minds, and Maybe Machines. Plenary Session 1 – *SRI International 2017 Technology and Consciousness Workshop Series*. Arlington, VA.

Zebrowski, R. (2017b). On the Possibility of Synthetic Phenomenology and Intersubjectivity. Phenomenology and Western Philosophical Perspectives Session – *SRI International 2017 Technology and Consciousness Workshop Series*. Arlington, VA.

Zebrowski, R. (2017c). Mind-Body Proliferation: The More We Learn, The Less We Know. Plenary Session 2 – *SRI International 2017 Technology and Consciousness Workshop Series*. Menlo Park, CA.