Trustworthy AI and the Logics of Intersectional Resistance

Bran Knowles
b.h.knowles1@lancaster.ac.uk
Lancaster University
Lancaster, UK

Jasmine Fledderjohann
j.fledderjohann@lancaster.ac.uk
Lancaster University
Lancaster, UK

John T. Richards
ajtr@us.ibm.com
TJ Watson Research Center, IBM
Yorktown Heights, New York, USA

Kush R. Varshney
krvarshn@us.ibm.com
TJ Watson Research Center, IBM
Yorktown Heights, New York, USA

ABSTRACT

Growing awareness of the capacity of AI to inflict harm has inspired efforts to delineate principles for 'trustworthy AI' and, from these, objective indicators of 'trustworthiness' for auditors and regulators. Such efforts run the risk of formalizing a distinctly privileged perspective on trustworthiness which is insensitive (or else indifferent) to the legitimate reasons for distrust held by marginalized people. By exploring a neglected *conative* element of trust, we broaden understandings of trust and trustworthiness to make sense of, and identify principles for responding productively to, distrust of ostensibly 'trustworthy' AI. Bringing social science scholarship into dialogue with AI criticism, we show that AI is being used to construct a digital underclass that is rhetorically labelled as 'undeserving', and highlight how this process fulfills functions for more privileged people and institutions. We argue that distrust of AI is warranted and healthy when the AI contributes to marginalization and structural violence, and that Trustworthy AI may fuel public resistance to the use of AI unless it addresses this dimension of untrustworthiness. To this end, we offer reformulations of core principles of Trustworthy AI—fairness, accountability, and transparency—that substantively address the deeper issues animating widespread public distrust of AI, including: *stewardship and care*, *openness and vulnerability*, and *humility and empowerment*. In light of legitimate reasons for distrust, we call on the field to to re-evaluate why the public would embrace the expansion of AI into all corners of society; in short, what makes it *worthy* of their trust.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Social and professional topics** → *Computing profession.*

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3593986

KEYWORDS

Trust, distrust, artificial intelligence, fairness, bias, inequality, intersectionality, accountability, transparency

ACM Reference Format:
Bran Knowles, Jasmine Fledderjohann, John T. Richards, and Kush R. Varshney. 2023. Trustworthy AI and the Logics of Intersectional Resistance. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3593986

1 INTRODUCTION

AI ethics is a rapidly growing area, within which sits a body of work self-described as 'Trustworthy AI'. In contrast to terms such as 'Responsible AI' or 'Ethical AI', the term 'Trustworthy AI' calls attention to popular concerns around *trust in AI*—concerns, perhaps, around who is trusting or distrusting, why this matters, and what might be done about it. And yet, these questions are marginal within Trustworthy AI; the overwhelming focus is the development of techniques and practices for a) making AI 'trustworthy' [57, 63, 64, 67, 94, 98] in accordance with various principles, and b) assessing it as such [2, 18, 75, 105]. The dearth of serious treatments of trust within the Trustworthy AI literature [55, 65] would seem to suggest that maybe trust in AI is simple: It is, rather uninterestingly, just a belief about AI's trustworthiness. But a doxastic (belief-based, or *cognitive*) account of trust seems too limiting for *trust in AI*. For one, the ability to form cognitive trust requires (among other things) that "the patterns that distinguish the model's correct and incorrect cases are available to the user" [45, emphasis removed]. Making this information available and interpretable to users has proven immensely challenging in the context of AI. The Trustworthy AI community is certainly aware of the potential for and dangers of incorrect cognition—the fact that dis/trust of AI can be "unwarranted" [36, 65, 90], leading to "inappropriate" reliance [59]—and the literature attempts to reconcile such instances where trust is not a perfect mirror of trustworthiness by exploring how *affect* moderates trust [42, 58, 60, 62, 91, 103]. But is it reasonable to treat trustworthiness as an objective quality of an AI system which is more or less recognizable depending on the correctness of cognition and the control of affect? Is it not possible that one person's distrust of AI is as "justified" [96] as another person's trust of the same AI, meaning that trustworthiness is at least partially subjective?

The case is clearer if we limit the discussion to socially applied AI. For example, systems that make decisions about how to allocate resources/opportunities across a population. If we take the definition of trust offered by Annette Baier, a "willingness to be or remain within [an entity's] power…and to give them discretionary powers in matters of concern to us," or more succinctly, "accept[ing] vulnerability to others" [6], then it is worth asking whether *some have more to gain and others more to lose in allowing themselves to be vulnerable to AI.*[1] To be clear, vulnerability arises from the potential for betrayal of trust, meaning simply that the trust ends up being shown to be misplaced. In this sense, to trust is to take a risk; but with risk comes the potential for reward [4]. Deciding to trust (to be vulnerable to possible betrayal of trust) allows a person to gain the benefits of reliance on a person or thing, while also gaining the opportunity to explore whether their trust proves to be well-placed and to better calibrate their future trusting. A person with greater resources and stability can more easily afford to take risks in trusting [4];[2] whereas in the absence of such privilege and stability, an individual may rightly perceive the risks of trust to be prohibitive, or at least to demand greater evidence of trustworthiness to justify their risk-taking (cf. trust motivation theory [97]).

The argument we set out in this paper is, first, that in order to understand distrust of AI by the public, we must make room for a *conative* element of trust, i.e. "judgments, decisions, intentions and resolutions which lead to a disposition to trust" [92]. This applies particularly to distrust of AI in the abstract sense, as conveyed by statements like "I don't trust AI". Our argument also helps to explain why there is variability in self-reported trust across the population [37, 49]. In contrast to two foundational papers in the field which see individual variance in one's "propensity" [69] or "predisposition" [59] to trust as a stable personality trait, we propose that conation is inherently context-dependent and, thus, subject to change if a material change in context is experienced. In §2 we provide an explanatory account of conative barriers to trust in AI, exploring how one's baseline perception of what one stands to lose or gain from AI is rooted both in personal experience and in collective experience—that is, how one's specific in-groups have been and continue to be harmed or benefited from inequitable social structures. Specifically, we show that AI is being used to construct a digital underclass that is rhetorically labelled as 'undeserving' so as to render the structural violence inflicted on them morally palatable. Drawing on Benjamin's language of "the racial logics of trust" [11] and "informed refusal" [12], we name the justified distrust of such AI by (multiply) marginalized individuals *the intersectional logics of resistance.*

Secondly, we argue that when distrust is understood as a legitimate response to a perceived threat, exploring this conative element is critical for taking seriously the legitimate concerns of those most likely to experience algorithmic harm. In §3 we critique current approaches to fairness, accountability, and transparency as responses to the conative issues animating public distrust of AI and offer initial musings on new aims to guide pursuits of 'trustworthiness'. This contribution helps attend to known shortcomings in the FAccT (and wider AI ethics) literature regarding consideration of AI impacts on marginalized groups [16].

Thirdly, we echo Kennedy's insight that "Seeing trust as a privilege enjoyed by majority groups might help us to resist the temptation to believe that more trust should be our goal" [49]. Ultimately, we argue that the conversation about whose trust is important needs to shift. To date, Trustworthy AI has focused on what makes AI 'trustworthy' to those who are highly motivated to trust AI, such as those wishing to integrate the technology into their business operations, those whose work requires that they interface with AI decision making tools, and policy makers who are keen to unlock AI's vast economic potential. And as noted by Lee and Rich [61], extant research disproportionately examines AI trust dynamics among populations who, unlike marginalized groups, have high trust in human decision makers and institutions. Recognizing that distrust of AI by marginalized individuals relates to reasonable distrust of the systems in which AI is increasingly implicated, the overriding moral obligation is to ensure that, for the most structurally disadvantaged in our society, the rewards of trusting AI exceed the risks of trusting AI.

2 INTERSECTIONAL LOGICS OF DISTRUST

Notions of what makes AI 'trustworthy' are typically oriented around high level principles developed by select experts (e.g. the European Commission's High-Level Expert Group on AI [79], and the Organisation for Economic Co-operation and Development's first expert group on AI [38] and now Working Party on Artificial Intelligence Governance [39]). While not the only approach to understanding trustworthiness in the context of AI, it is a prominent and influential view, rooted in the authority of a high-level expert commission. We contend that these experts will have a particular *viewpoint* which can be characterized as a view-from-the-top: not only is this view largely ignorant of *being* from-the-top, but from this vantage the benefits of AI are clear and the harms of AI are abstract, largely separate from lived experience[3] but a potential threat to an ambitious technology agenda. This view-from-the-top has thus informed highly technocentric principles for 'trustworthiness' [65] which, as we will show, obscure and facilitate structural violence by shifting attention away from how the very foundations of AI are inherently extractive and prone to reproducing and, at the same time, amplifying extant inequitable social structures through the logic of categorization and simplification (see, for example, [21, 71]). Instead, this framing of trustworthiness gives the impression that AI has incalculable potential to increase efficiency and improve lives, with only a few small tweaks from good faith actors being needed to safeguard against the occasional unintentional shortcoming, e.g. bias. Framing AI as inherently objective and trustworthy in this way therefore makes it more likely to inflict the kind of harms that would promote *warranted* distrust.

In the next subsections we explore the *intersectional logics of distrust*, i.e. the reasons individuals experiencing inequality and marginalization might distrust AI. We contend that AI is particularly prone to cause harm to marginalized groups. And, building from legal scholar and rights activist Kimberlé Crenshaw's concept

[1] Jason D'Cruz, personal correspondence.
[2] Also: Mark Alfano, Facebook post, October 24, 2022.

[3] D'Ignazio and Klein refer to this phenomenon as the "privilege hazard" [26, p. 29], or the inability of those at the top to recognize oppression of those below.

of intersectionality [22], we further assert that multiply marginalized people are subject to unique harms which can exceed the additive harms inflicted on each group to which multiply marginalized people belong.[4] We argue that existing forms of marginalization have been used to construct, monitor, monetize, and penalize a *digital underclass*, whose individual and collective experiences of structural violence outside of the digital realm are reproduced and compounded by structural violence within this realm [13, 19, 32, 71]. Longstanding, harmful narratives that frame marginalized people as undeserving (of social protection, of restorative justice, and of empathy) have been directly transferred into the discourse around digital technologies and AI, providing rhetorical cover for harms inflicted on this digital underclass. We bring social science scholarship into dialogue with AI criticism by exploring the positive functions of the construction and maintenance of this 'undeserving' digital underclass for the people and institutions which marginalize people We place this dialogue within a broader discussion on structural violence and trust in AI. As phenomena can simultaneously perform different functions for different groups, we also consider how intersectionality can help us to identify for whom these positive functions operate. We look at the unique, additive harms inflicted upon the digital underclass, and we consider implications of these harms for trust.

2.1 Construction of an undeserving digital underclass

At the core of our argument is the assertion that distrust in social systems is a reasonable, warranted response to marginalization. In this section, we consider how construction of an 'undeserving' digital underclass systematically marginalizes some groups, and we explore the functions this act of marginalization serves for others.

In his 1994 work "Positive Functions of the Undeserving Poor", sociologist Herbert Gans [41] articulates the myriad positive functions that the existence of an 'undeserving' economic underclass fulfills for society writ large. He acknowledges that poverty clearly has a range of well-understood negative functions,[5] disproportionately (but not exclusively) experienced by this economic underclass. However, he adds to this that the rhetoric of undeservingness (e.g. popular misconceptions that poor people are lazy, are dependent on welfare, engage in unsanctioned behavior such as drug use and criminalised behavior) is utilized to render palatable the often invisible harms inflicted on this group. This undeservingness narrative implies that poor people do not merit the social protection that could help them to overcome structural economic barriers, and so the narrative facilitates perpetuation of the stratified social order as morally justifiable.

Positive functions help to explain why phenomena persist. Here, we explore a few of the positive functions of the 'undeserving' poor Gans outlines. One positive function is to provide a scapegoat on whom individuals and institutions can blame social problems, thereby alleviating the need to address social problems. Another important function, which benefits employers in particular, is that the undeserving poor provide a reserve labor pool who can be excluded until they are needed to fill labor deficits; simultaneously, they serve as a pool of 'hypothetical workers', whose very existence serves to drive down wages and suppress union organizing activity by implying that employees are ultimately replaceable. This function is closely linked to another: Some people who are locked out of the labor market may turn to the manufacture and sale of illegal goods, which more privileged groups can then acquire. Additionally, because they are rhetorically framed as poorly socialized or even dangerous, the undeserving poor create jobs for the more privileged, such as social workers, judges, prison guards, teachers, journalists, and social scientists. Among other functions, the undeserving poor also provide pop culture villains, which is both a marketable (profitable) narrative for producers of television, films, and other media, and also serves as a self-reinforcement mechanism. Perhaps most bleakly, Gans explains that the ill health and premature mortality of the undeserving poor, which has been widely accepted by society by virtue of their 'undeservingness', reduces competition for economic resources among the remaining population.

Building from Gans, we argue here that popular undeservingness narratives are being both reproduced and amplified by AI. Much like analog inequities, this process both allows the 'deserving' to reap the benefits of subjugating an underclass and alleviates the imperative to address algorithmically-enhanced inequities and the logics that drive them. Indeed, through the narrative power of undeservingness, inequities can be recast as *inequalities* by suggesting that these inequalities are just. In short, the notion of undeservingness is powerfully deployed in ways that obfuscate the structural violence of AI by framing it as merited. This in itself is not new territory for AI ethics; yet while much attention has rightly been given to the negative functions of an undeserving digital underclass for the field of AI, less attention has been given to the positive functions.

Without aiming to provide an exhaustive list, we can identify some of these positive functions.[6] We highlight that a single phenomenon can have positive, negative, and neutral functions, and, linking to Crenshaw [22], suggest that whether one experiences the negative functions of the phenomenon is intimately linked to one's intersectional position within the social structure; multiply marginalized groups, who tend to be relegated to the digital underclass, are most likely to experience negative functions. Within the context of Trustworthy AI, we argue that an 'undeserving' digital underclass has been created—a group whose marginalization in

[4] We use the language of marginalization to draw attention to structural processes that systematically devalue and disadvantage some groups. It is not an inherent characteristic of people but rather their *systematic exclusion* that generates inequities and harms. We use the term to capture vast systemic processes that exclude and harm people on the basis of disability, racial and ethnic categorizations, gender, age, sexual orientation, migration status, and other markers for structural violence, but also recognize that marginalization is contextual. Some groups will be more marginalized in specific contexts than others, and some individuals will identify as (or be externally identified as) belonging to multiple marginalized groups, resulting in a compounding of marginalization. Our language choice here aims to capture that marginalization is a systematic *process* of structural violence.

[5] Some scholars may prefer the language of functions versus dysfunctions in deference to Robert Merton [73]; we employ the language of positive versus negative functions because we engage with Merton only indirectly through Gans here, but recognize Merton's work was seminal for understanding functions.

[6] To be clear, we do not use the language of 'positive functions' to defend the construction of the digital underclass in the AI sphere. Just as Gans explains that the existence of positive functions does not *justify* undeservingness narratives but rather explains the persistence of such narratives, we aim to highlight how the digital underclass selectively benefits—that is, serves positive functions for—specific actors in the AI sphere. A 'positive' function does not mean a function that is morally good.

analog spheres is mirrored and compounded in the digital realm, resulting in structural violence against the many and considerable benefits for the few [7]. In the subsections that follow, we examine some of the negative functions that have been rightly identified as pernicious in previous literature, and we illustratively identify some of the ways that these same functions that are negative for some groups are positive for others (disproportionately for the privileged).

> *An explanatory note:* We do not believe our argument requires us to determine whether positive functions are intentionally produced. Systems need not be designed with the intention to harm in order for them to cause harm. Those who benefit from inequitable systems risk being complicit in structural violence unless they actively work to understand and dismantle oppression.

2.2 Widening inequality and the functions of the undeserving digital underclass

Tracing the functions of the digital underclass requires active attention to the beneficiaries of extant (digital) social structures. Who benefits from the construction and maintenance of the digital underclass, and how? An important clue in this context is institutional alignment. When we consider some of the institutions eagerly utilizing AI and examine what that AI is being deployed for, the construction of the digital underclass and the positive functions come into sharper focus. For example, AI systems are being avidly deployed in criminal justice systems around the globe not only to surveil populations, but also to make judgements about criminal risks and justify actions such as arrests and deportations [21]. On the surface, if the aim of the criminal justice system as an institution is to enforce the law, AI systems that promise to save labor costs and improve efficiency seem from an institutional perspective to be a reasonable measure for making the system more effective. However, this stated goal is inherently in tension with the safety and well-being of marginalized groups. Multiply marginalized people are substantially more likely to be discriminated against in the very language and substantive focus of laws themselves, and are disproportionately targeted for surveillance and violence by the police, falsely convicted, and sentenced for lengthy prison terms (see for example [3, 23, 24]).[8]

Michelle Alexander's work on the "New Jim Crow" identifies how the legal system has been utilized in the United States to remove Black people from society and into prisons, thereby legally stripping them of their rights to access social services and institutions (e.g. education, housing, employment, democratic enfranchisement, social protection) . These are the same services and institutions that were systematically denied to Black people through Jim Crow laws prior to the Civil Rights Movement. In short, while Jim Crow laws have been struck down as discriminatory, mass incarceration provides a sufficiently opaque legal means of reaching the same ends. Extending Alexander's work to the context of digital inequities, Ruha Benjamin [11] introduces the concept of the "New Jim Code" to explain how new technology is being used to reframe the perpetuation of structural inequities as objective and/or progressive under the false premise that such technologies are neutral and, therefore, not discriminatory. AI, then, can be better understood as a tool not for fairly and effectively enforcing the law (as the institutional rhetoric would suggest), but rather for giving rhetorical cover while also efficiently amplifying existing patterns of structural violence in the criminal justice system (as well, of course, as in other institutions—a point which we explore further below). Moreover, McQuillan [71] asserts that AI treats structural characteristics as ultimately creating a system of "self-reinforcing social profiling." Importantly, while a growing body of literature on mass incarceration focuses on the United States, we reference this here as an illustrative example; the broad pattern of legally sanctioned marginalization and structural violence, which is being both rhetorically and practically supported by supposedly objective emerging technologies, is certainly not unique to that country.

For a privileged person deemed low-risk by algorithmic classification, as well as for the institutions which claim to uphold social order, AI serves a positive function: It decreases the moral panic towards the generalised 'criminal' by promising that this undeserving underclass will be effectively identified and isolated from society, creating the feeling of safety and security. Another positive function which serves broad swathes of the privileged public is norm reinforcement: AI algorithms can heighten the efficiency of identifying (and penalizing) violations of social norms, and likewise of rewarding adherents. On the other hand, for someone who is more likely to be algorithmically classified as high-risk on the basis of their skin tone and other physical characteristics, name, religion, neurodivergence, gender identity, and/or any other markers selected for marginalization, AI carries the very real risk of relegation to the undeserving digital underclass (a negative function to say the least). How or why a rational actor would trust a system that more effectively targets them for structural violence in this manner is unclear. Understanding the positive functions for the privileged helps to highlight why some groups may have "motivated cognition" [97] in regarding AI as generally trustworthy as much as the negative functions explain this motivated cognition in the other direction.

Recent applications of AI in the retrenchment of social protection schemes similarly follow a pattern of marketing structural violence against marginalized people as progressive and morally palatable by invoking AI's alleged objectivity. Looking, for example, at the Aid to Families with Dependent Children / Temporary Assistance for Needy Families program in the United States, commonly referred to simply as 'welfare', Black feminist scholars [85–87] have been calling out the intersectional pattern of discrimination inherent in the abrogation of this social protection for decades. Specifically, while there is an overarching stigma (rooted in moral judgements

[7] Although analog marginalization may create the conditions for warranted distrust in digital systems, we do not propose a 1:1 mapping exercise, where distrust in digital systems is directly or solely rooted in analog harms and associated warranted distrust. Our point is neither that distrust is *only* warranted where it is rooted in analog harms, nor that distrust in digital harm must be *proportional* to distrust in analog systems. Rather, we argue that distrust of marginalizing systems is warranted, and that replication and amplification of analog marginalization by digital technologies is a process that may rightly generate (further) distrust.

[8] We note for now that the establishment of a regulatory ecosystem for Trustworthy AI is hardly a comfort to those whose experience of the legal sphere is as a mechanism for solidifying inequity.

about single parenthood and stereotypes about laziness and unwillingness of recipients to work[9]) attached to accessing this program, widespread falsehoods about so-called 'Welfare Queens', who are presumed to have (too many) children in an effort to scam taxpayers and avoid real work, have been used to justify withdrawal of social protection in an act of structural violence. Crucially, the Welfare Queen is racialized as Black, and racist tropes intersect with class and gender norms to make parenting in socially approved ways uniquely impossible for low-income Black women.

Linking back to Benjamin's [11] "New Jim Code," what is new in the long history of vilification of (multiply) marginalized mothers is the use of AI to increase the efficiency of marginalization, and use of the associated language of objectivity to render this structural violence of social exclusion morally palatable. As Eubanks explains, algorithmic decision making tools "manage the individual poor" through practices of profiling, surveillance and punishment, "in order to escape our shared responsibility for eradicating poverty" [32]. Benthall [14] adds that AI systems that allocate resources tend to either reify racial categories or reify "racialized social inequality by no longer measuring systemic inequality." While it is of course true that poor white men, for example, are also impacted by classist AI, an intersectional view of these systems shows that the feminization of poverty and the gendered patterns of single parenthood and receipt of social protection mean women are disproportionately likely to be affected by some of these class-based acts of structural violence, and even more so for multiply marginalized women. And, as with widening of the carceral net, stigma, marginalization, and the growing role of AI in inflicting this form of structural violence through the construction of an undeserving digital underclass is by no means unique to the United States. For example, in the Netherlands, the SyRI (Systeem Risico Indicatie) algorithm was applied to a disturbingly wide range of government databases to identify individuals who may be 'guilty' of committing benefits fraud [101]. The algorithm, which was applied with no democratic oversight, specifically targeted deprived neighborhoods.

While tremendously harmful for recipients of social protection, stigmatizing stereotypes and the rhetoric of undeservingness again facilitates the positive function of norm reinforcement for institutions such as the family and the economy. Norm reinforcement provided by social protection recipients occurs in analog space, but has been supercharged by AI, under the presumption that the 'undeserving', who supposedly abuse the system, can be precisely and efficiently weeded out from among the 'deserving' few who are deemed to be in legitimate receipt of social protection. In this skewed view, the undeserving digital underclass are rendered increasingly incapable of abusing taxpayer money because AI will identify and penalize any malfeasance. Indeed, in direct parallel to Gans's functions, the undeserving digital underclass can be scapegoated, alleviating the need to address social problems.

The undeserving digital underclass is particularly useful in the context of maintaining the inequitable neoliberal state. The narrative that state divestment in the form of austerity and further opening of 'free' markets is the solution for a failing state is propagated by neoliberal politicians, who claim the state's financial woes result from irresponsible spending on social protection [19]. By rendering inequity and decimation of social protection systems morally palatable, and by constructing a digital underclass that can be scapegoated and penalized for the failures of the state, AI can be deployed by neoliberal politicians to give the impression that efficient action is finally being taken to redress reckless spending on social protection. As McQuillan states, "AI represents a technological shift in the framework of society that will amplify austerity while enabling authoritarian politics" [71, p. 10]. We add that, as a rhetorical bonus, the utilization of AI systems developed through private enterprise and apparent competition among companies selling AI products in the free market can be touted as neoliberal government adhering to its own free market principles.

That AI serves positive functions under a free market state system has clear implications for the economy as an institution. One positive function of the digital underclass under the principles of a capitalist economic system is that the underclass can be exploited to generate revenue. As explored by Safiya Umoja Noble, this operates directly in web search algorithms which, in their ostensible role as neutral purveyors of relevant information, will offer up racist and sexist results because doing so is, as the algorithm has learned, profitable [77]. There is also an indirect pathway linked to revenue generation through the impacts of AI on labor market dynamics. Specifically, the underlying human labor needed to gather and code the data at the heart of AI generates a wide array of jobs. Importantly, however, these are often high quantity but low quality precarious jobs; AI can harness the narrative of 'job creation', which is well-measured through existing metrics, but with less accountability for the quality of these jobs, which are not systematically measured and monitored in the same way.

We therefore include in our conception of the undeserving digital underclass not only those who are harmed by implementation of algorithms, but also those who are harmed in their construction, such as 'low-skilled' workers who are devalued for supposedly insufficiently leveling up their skills and qualifications. These essential workers are left to take on emotionally taxing, poorly remunerated, and unrewarding labor with limited job security and prospects for promotion (see for example [21]). The creation of AI training data by low-level employees can generate billions of dollars of wealth, little of which ever flows to their benefit. Much as Gans argued that the undeserving poor provide a reserve labor force and also drive down wages and suppress union activity, so too does the digital underclass fill this role, ironically accepting precarious and low wage work to build the very AI systems that are being used to make them feel replaceable. Laborers also become enfolded in the digital underclass through increasing forms of workplace surveillance, which create a digital panopticon in which workers are assumed to be lazy and ineffective, stealing company time if they are not directly monitored. This highlights how porous the boundaries of the undeserving digital underclass are—marginalization creates a gradient rather than a binary set of categories, so that multiply marginalized people experience the most consistent and severe consequences of marginalization, including bearing the brunt of undeservingness narratives, while less marginalized groups can contextually experience both positive and negative functions.

[9]As social reproduction theory [15, 35] asserts, the notion that caring work to birth, feed, and raise children, is not 'work' is in and of itself a misogynistic and highly problematic notion that systematically devalues women's immense contributions to economic systems.

As the examples above show, the undeserving digital underclass performs a positive function by fostering faith in existing systems: Society can have confidence in a system that perpetuates and amplifies stark inequities because an 'objective' algorithm has identified the 'undeserving'. Under this logic, societal norms are assumed to be beyond reproach because seemingly unbiased systems have confirmed the legitimacy of the existing, inequitable social order. Linked to this, the digital underclass absolves the privileged of the need to examine the morality of phenomena such as the abrogation of social protection and the widening of the carceral net because the assumption is that these systems have efficiently identified those who are inherently undeserving of support and deserving of punishment; absent of messy human error introducing false positives, moral concerns around issuing lengthy prison sentences or withdrawing social support from someone in need are rendered moot.[10] Implicitly, *the privileged public can trust unbiased AI to get it right.* Unfortunately, in reality, examples of deeply biased AI abound. Moreover, the very notion that *anyone* is undeserving and must be identified and penalized illustrates a fundamental problem with the entire foundation: Algorithms are being constructed by human beings, who are subject to socially developed biases; AI is not created in a vacuum, but instead hinges on, among other (biased) inputs, human-driven collection and categorization of data. Rather than operating beyond human bias, AI can very efficiently achieve deeply biased human goals. Notably, the marginalized public have no cause for optimism regarding AI's capacity to render unbiased decisions; and in fact they need not have direct experience of betrayal by biased AI to be wary of it. Lived experience of institutional callousness within the inequitable system these technologies are designed to optimize is sufficient cause for distrust.

3 IMPLICATIONS FOR FACCT

Having explored distrust of AI by marginalized individuals (admittedly, saying nothing of the prevalence of these attitudes, which will be the focus of future work), we now proceed to identify shortcomings of common strategies for promoting public trust in AI. Specifically, we argue that moral commitments that would ostensibly engender trust have been operationalized in ways that may, perversely, contribute to public distrust of AI. To correct this, we propose alternative framings of accountability, transparency, and fairness that are more likely to promote well-placed trust by disrupting the positive functions that incentivize harm—in short, materially and proactively addressing the harms upon which contextual judgments regarding AI's trustworthiness are based. What follows is exploratory, and offered as food for thought. Our recommendations below build on insights generated through examinations of power dynamics which negatively shape perceptions of AI [1, 7, 82, 88, 93], and have deep resonance with feminist approaches to AI (e.g. [52, 95]) insofar as they call for appreciating intersectionality, actively dismantling structural privilege and oppression, and engaging with marginalized people to give voice to their experience (see also: [26]).

[10] Green [43] makes this point as it relates to (criminal) risk assessment algorithms "legitimizing the criminal justice system's structural racism."

3.1 Reconsidering 'Accountability'

It has been noted elsewhere the prevalence of instrumental arguments for the importance of public trust [55], and that such arguments hinge on the false premise that the public has agency to rely only on AI they trust [54]. To this we add that despite rhetorical overtures to 'the public', as if this body is homogeneous, instrumental valuations of trust do not require that those developing AI systems care at all about earning the trust of people who are structurally disadvantaged (see also: [100]) because they have the least power to challenge AI they deem 'untrustworthy'; in fact, as explored in §2.2, there may be more to gain from their marginalization, and thus from AI that is 'untrustworthy' in this fundamental respect. In what way, then, is the AI accountable to 'the public'?

In the context of Trustworthy AI, accountability is a process for assessing conformity with a set of agreed principles and objectives. Importantly, it is one class of experts communicating with another class of experts regarding this conformity. The aims of accountability include *traceability* [56], and perhaps *verifiability* [27], and these offer important safeguards (albeit of a particular order defined by an elite class of individuals [44]). But "following prescribed procedures and requirements" in and of itself should not be construed as the "proper aim" of Trustworthy AI (paraphrasing [81]). A checklist of technical parameters and processes that plausibly relate to a form of trustworthiness provides a mechanism for signaling that ethical concerns are being taken seriously without requiring a deeper and more nuanced dive into the subtleties of what makes AI worthy or unworthy of trust. So in this sense, adhering to typical accountability practices can obscure what would actually be needed to engender public trust in AI (cf. [20]).

In articulating a 'proper aim', it is worth recalling that the core feature of trust relationships is the entrusting of one's vulnerability[11] to another [5, 89]. The act of entrusting generates a responsibility on the part of the trustee to not betray the trustor's vulnerability, and being trustworthy means taking that responsibility seriously. As D'Cruz notes, "If you think a person is indifferent to the fact of your vulnerability, or that a person is hostile to you, then you will distrust them across multiple domains of interaction" [28]. This simultaneously offers an explanation for potentially quite diffuse distrust of AI[12] by marginalized individuals and points to a new moral aim to guide accountability efforts: *stewardship*, i.e. the careful management of the vulnerability entrusted to one's care.

An important effect of centering vulnerability in this way is to direct accountability away from the regulatory apparatus or an organization's internally defined processes and metrics and toward the entrusting parties themselves—a theme we notice emerging within accountability literature [20, 48, 74, 104]. Paradigmatic features of trust include, in addition to the entrusting of one's vulnerability, that the trustor is optimistic that the trustee is competent and committed to doing what they are trusted to do [25]—in the best case, that a trustor recognizes that the trustee *cares* about them, above and beyond the repercussions they might face for failing as

[11] Again, we use the term vulnerability as defined in the literature on trust; it is not a claim of disempowerment nor restricted agency. An empowered agent may choose whether or not to open themselves to the vulnerability entailed in trusting.
[12] For example, distrusting AI "to make any decisions" about one's life [9].

stewards.[13] Taking inspiration from care ethics literature (e.g. Jean Watson's theory of human caring [102]), a *carative approach*[14] to AI would de-center statistical modelling within the AI development lifecycle to make room for core practices of accompaniment of and caring for the most vulnerable. Development teams might benefit, therefore, from a facilitated process to help them overcome their "meta-blindness", or failure to see what they are not seeing from others' perspectives [89, referencing [72]], not just through collaboration within more diverse teams [34], but by those teams "intentionally and attentively placing themselves in situations in which they will experience epistemic friction" [89] throughout the AI development lifecycle (e.g. [52]).[15] At a minimum, a carative development process would 1) start with real-world problems as experienced by the most vulnerable; 2) listen to them to understand their values, concerns, and constraints; 3) meet them where they are and work toward a solution to their problems all the way to the end; and 4) conduct a grounded assessment of the entire solution by meaningfully engaging with the affected communities both before and after deployment.

3.2 Reconsidering 'Transparency'

In recommending the above, we are well aware of the practical barriers (cost, resources) faced within AI development. Finding, let alone bringing into the requirements and testing process, the range of potentially affected parties is a significant challenge, and not everyone involved in developing AI has the skill set for engaging in these explorations. There may be ways of reducing costs, e.g. creating shareable training datasets that have been thoughtfully curated and establishing diverse subject pools[16] that could serve as expert advisors across a range of AI projects. But there is no getting around the fact that AI development teams will never be able to understand, anticipate, and mitigate harms from all perspectives. This being the case, a powerful gesture of trustworthiness would be to be *open* about one's evolving understanding of harms and how (and by whom) this has been informed—to first reflect on, and then lay bare for public critique, how those developing AI see the world. To quote Onara O'Neil, "Speaking truthfully does not damage trust, it creates a climate for trust" [80], as it signals a determination not to deceive (for the AI not to be trusted more than it deserves to be) and a genuine striving toward trustworthiness.

This open posture transforms transparency from backward-facing (what one has done in the development process) to forward-facing (what one needs to find out) while also creating a channel for receiving continuous input from the public.[17] In doing so, it corrects what has never quite worked regarding transparency's relationship to accountability: The fact is, the lay public is characteristically unable to evaluate AI's trustworthiness if offered the sort of evidence we know how to produce [55], let alone leverage such evidence for greater control over the AI [66] or toward remediation of harm [17, 99]. So despite being potentially vital to the work of other parties in managing compliance, transparency as mere *disclosure* [78] can inflict "epistemic injustice" by "rendering [the public] unable to challenge injustice" [71, referencing [40]]. AI that is worthy of public trust would allow non-specialists to meaningfully engage in conversations about its appropriateness, and provide real opportunities for the public to reject AI deemed inappropriate—i.e. "encouraging genuine dialogue" [50] through mechanisms such as "People's Councils" [71]. There is a promising opening for regulations and standards to begin shifting AI development toward a more open and productive relationship with the real harms it might be causing (and, indeed, the real opportunities for promoting the general welfare it may not be considering). By stipulating the inclusion of a diverse range of stakeholders (see [79] Chapter II, Secion 1.5) and by suggesting a range of elicitation techniques to explore their perspectives, situations for true discovery might arise.

We are also struck that openness becomes especially transformative when, in making oneself open, one allows oneself to be *vulnerable* to the other party. As noted by Baier [5], there are many kinds of trust relationships, some more "morally decent" than others, such as those rooted in mutual respect and symmetry (in power and intimacy). Inspired by the work of Nagar [76], Arif and Os write: "Vulnerability becomes radical when we are critically open about it and collectively surrender to it to forge knowledge-making relationships that are more rooted in solidarity" [4]. We note that AI seeks to minimize vulnerability of the deploying organization (e.g., loss of income due to a loan defaulting) [53], while increasing vulnerability of those subject to algorithmic decision making (as explored in §2). So what might it mean for an AI developing organization to become more vulnerable, better balancing the vulnerability that is now borne primarily by those subject to AI? The answer seems to lie in becoming more aware of the vulnerabilities the organization actually faces (e.g., the loss of business or reputational harm that can result from denial of opportunity) and being transparent about these vulnerabilities. This would also open the organization to considering how to reduce these harms which could also be reflected in their public stance.

[13] Trust scholar Annette Baier writes, "the best reason for thinking that one's own good is also a common good is being loved" [5].

[14] This terminology, 'carative AI', has the added advantage of explicitly valuing the work of caring as a challenge to the ways that AI is complicit in systematic devaluation and marginalization of those typically doing caring work (see Social Reproduction Theory, e.g. [15]).

[15] Important engagement work by independent research bodies such as The Ada Lovelace Institute is giving voice to manifold perspectives and involving diverse publics in defining what constitutes 'trustworthy AI', but more of this work needs to be done by those developing and deploying AI.

[16] We caution that simply establishing diverse subject pools (and diverse tech workplaces) cannot be considered sufficient. As sociologist Steve Epstein [30] argues in the context of medical inclusion, that exclusion can cause harm is well-established, but there are in fact a myriad of ways *inclusion* can cause harm also. For example, women's heart attack symptoms, which differ from those of men, have been poorly understood because women were historically excluded from research in part due to their 'vulnerability', particularly during pregnancy. On the other hand, Black people have long been included in medical research in tremendously harmful ways, with their bodies being used as testing grounds for deeply unethical medical experimentation (the infamous Tuskegee syphilis study being only one of countless examples). This is no different in many ways than how the undeserving digital underclass are being included in technology—as products and subjects, but not as legitimate stakeholders. How and why people are included is vitally important in determining whether inclusion reduces inequity. It is essential that people be empowered to set the terms of their inclusion, and that their inputs be genuinely valued. Otherwise, not only will inclusion be tokenistic and ultimately ineffective, but it will also not be any more worthy of trust than the current system.

[17] This openness to input is similar to the stance adopted by the Lean Startup methodology [84] and agile software development [31] more generally. Rather than follow a prescribed path dictated by requirements, these approaches progress through a series of small experiments (cast as Minimal Viable Products in the case of Lean Startup) which progressively reveal answers to what is most unknown about where true value might lie.

3.3 Reconsidering 'Fairness'

The dominant concerns within AI fairness discourse are the measurement and mitigation of bias. Predictions or prescribed outcomes by an AI are considered unfair if different groups (socially constructed categories relating to protected attributes such as race or gender) experience different rates of false positives, false negatives, or favorable/unfavorable decisions; or if individuals who are similar (e.g. similarly qualified) receive different outcomes. Reducing statistical bias often causes a reduction in predictive accuracy, at least with respect to the potentially biased training data (and associated set aside test data) that might have given rise to the bias in the first place. So this might be seen as further evidence of the moral commitment of the party removing that bias to avoid discrimination. Documenting this bias removal in a way that satisfies future audits may also reduce the likelihood of penalties (once guidelines spawn enforceable standards). So this might be viewed as a win all around, except for the fact that systemic discrimination cannot be reduced to the sort of statistical measures offered by the variety of bias toolkits now available (for example [10]). In the bigger picture, what makes AI unfair to marginalized groups is that they bestow an illusory objectivity upon categorizations that reinforce existing social hierarchies and attendant inequities, thus allowing the scaling up and simultaneous sterilization of slow violence done to the supposedly 'undeserving'.

We propose that remedies to the above require the adoption of different moral aims to guide 'fair' AI in place of non-discrimination. Firstly, we suggest that AI systems should manifest *humility* regarding their ability to discern the 'undeserving' from the 'deserving' (cf. [53]). Indeed, the very notion of un/deservingness is inherently discriminatory and rooted in human bias because it implies not all humans are worthy of the same positive outcomes—that our characteristics and/or behaviors can be used to justify restricting our access to some social goods. No matter how well-crafted, no algorithm can ever enforce such a biased premise in an unbiased way. Most socially applied AI is premised on the imperfect assumption that a person's past behavior (captured by a limited set of features) is meaningfully predictive of their future behavior. Hypothetically, if an institution were truly interested in fairly allocating resources to individuals, rather than using past behavior as a justification for withholding resources, they might instead seek out information about the context of a person's past behavior to understand what structural constraints may have shaped this past behavior i.e. "morally excusing conditions" [29]. Instabilities created by poverty, precariousness, heightened surveillance, and so forth, create challenges for marginalized individuals in demonstrating seemingly virtuous behavior, such as compliance [33]. Furthermore, what is interpreted as indicative of deservingness will tend to align with norms established by the dominant group. Humility, therefore, implies continually looking to expand the range and types of evidence being used as model features to "allow people to show themselves more fully" [53] while improving techniques for determining causality [51]; involving humans within the decision making process in ways that actually allow them to exercise their unique capacity for empathy [29]; but most importantly, remaining skeptical of what AI can really discern regarding a person's character [8, 29] given the strong influence of situational factors [53].

Even more radically, what if AI systems aimed to *empower* individuals in ways that, as much as possible, everyone is given an equal chance to succeed? As noted by Dan McQuillan, "As with so much of socially applied machine learning, the algorithms simply end up identifying people with complex needs, but in a way that amplifies their abandonment" [71, p. 80]. What if this capability to identify people's needs was subverted and put to good use? Instead of being seen as a tool for differentiating people as landing above or below a decision threshold, AI that embodies a more substantive commitment to fairness would instead be used to assist society in better targeting the right opportunities and resources to the right people so that everyone receives the support they need. Linking back to the roles of institutions in controlling the allocation of resources across society—which, as argued in §2, is interpreted within a neoliberal austerity frame as a mandate to more efficiently target resources to the 'deserving'—for those receiving little, it is natural to think that any enhancement of an institution's power is coupled with a further reduction in resources allocated to the marginalized. This also means, however, that more efficient distribution of resources to those in need might, if presented effectively, increase trust not only in the AI used to deliver these benefits but also in the institutions utilizing them.

4 CONCLUSION

> *"Resistance is not a force to fear: it is a powerful signal…Harnessed well, public resistance can help shine a light on what must be improved, weed out AI 'snake oil', define what is socially acceptable, and help a more responsible AI industry flourish" (Aidan Peppin, in [47]).*

There is a risk of Trustworthy AI being used for the purposes of 'trust washing' unless the efforts pursued under its heading meaningfully relate to the concerns underlying expressions of distrust—especially if the evidence of trustworthiness offered is compliance with guidelines and regulations which are themselves only weakly reflecting (as opposed to actively dismantling) structural violence. In focusing on conative trust we are not meaning to suggest that it is productive to consider conation in isolation from cognition and/or affect. These are interlinked, as Baier herself strongly argued [6] and trust motivation theory reiterates [97]. Rather, our point is that ignoring the relational-motivational aspects of trust—following the same pattern of oversimplification of the human experience which underlies much of the harm inflicted by machines [8, 21, 70, 71]—is a convenient way to performatively vet AI's trustworthiness while sidestepping the matter of whether people *actually* trust AI, or indeed whether they *should*.

The public is frequently polled by researchers and marketers regarding their level of trust in AI, but the rhetorical intention of their responses is often unclear. Distrust may indicate concerns about AI's reliability, concerns about the trustworthiness of the deploying organization, concerns about the efficacy of regulatory safeguards, or any number of moral concerns about AI [55]. With this paper, we have provided a plausible account of some of the concerns that the public might be conveying with distrust, particularly when a person's distrust of AI is categorical: they are "seeing through" to how AI is "connected to larger systems of institutionalized oppression" [83]. Part of our reason for doing so is to distinguish this kind of

distrust from non-reliance, especially given that much more of the focus of Trustworthy AI is on matters of 'reliance' and 'reliability'. As D'Cruz notes, "To apply the label 'untrustworthy' is to impugn a person's moral character. But to recognize that someone is not to be relied on in a particular domain need not have any moral valence at all" [28]. The distrust we have been describing *is* about the moral failings of AI. It is a harsh indictment, indeed! We risk eviscerating the important matters being implicated by distrust by conflating them with mundane matters of reliability; and we misunderstand their moral weight by failing to distinguish distrust (an active stance of opposition) from non-trust (a lack of reason to trust).

This paper does not provide scope for a full inventory of the violence that might provoke public distrust. But even this brief reflection reveals distrust as not merely logical, but *healthy*. As noted by Matthes, "cultivating a healthy distrust, particularly of elected representatives, is constitutive of a well-functioning democracy" [68]; likewise, we propose, a healthy distrust of AI serves as a check on the more oppressive (even sometimes "fascist" [71]) tendencies of socially applied AI. The almost universally held goal of Trustworthy AI, to promote trust in AI, rather misses the point [49]—if, that is, it is taken as equivalent to reducing distrust of AI. Echoing the epigraph above, distrust as moral resistance is not what we should be eliminating, but rather what we should be keenly focused on understanding because it signals a social justice issue that needs attending to.[18] Distrust in AI may wane as the field responds to the moral objections of publics; then again, distrust may not depreciate in absolute terms, but change in quality, reflecting a forever changing landscape of social justice concerns. What is important is not its waning but that it has inspired better, more socially just, AI.

Our discussion of the logics of intersectional resistance is currently speculative. Empirical research exploring interactions between various axes of social marginalization, experiences of AI harms, and distrust of AI is greatly needed to inform this continual betterment of AI. And although we have argued that the elimination of distrust is not a meaningful (or achievable) goal, tracking changes in attitudes to AI over time for particular sub-groups of the public would serve as a valuable indicator of whether and how advancements in AI are affecting a changing distribution of benefits and harms across society. Ultimately, by opening up conversations around AI 'trustworthiness' to considerations of their complex entanglements with inequitable systems and structures, we may realize the much more meaningful aim of substantively addressing those inequities.

ACKNOWLEDGMENTS

This work is partially funded by the ESRC funded grant BIAS: Responsible AI for Labour Market Equality (ES/T012382/1) and by the EPSRC funded grant Equity for the Older: Beyond Digital Access (EP/W025337/1). We would like to thank our colleague Jason D'Cruz for helping to shape our thinking in this space, and the anonymous reviewers who helped us improve the work.

[18]We would be remiss if we did not acknowledge being inspired by a particular passage in a paper by Johnson and Melnikov [46], cited by [12, 49]: "the problem of distrusting citizens should be recast or reformulated as an issue of social justice."

REFERENCES

[1] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. 2021. Narratives and Counternarratives on Data Sharing in Africa. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 329–341.
[2] Mohit Kumar Ahuja, Mohamed-Bachir Belaid, Pierre Bernabé, Mathieu Collet, Arnaud Gotlieb, Chhagan Lal, Dusica Marijan, Sagar Sen, Aizaz Sharif, and Helge Spieker. 2020. Opening the Software Engineering Toolbox for the Assessment of Trustworthy AI. *arXiv preprint arXiv:2007.07768* (2020).
[3] Michelle Alexander. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colourblindness.* The New Press.
[4] Ahmer Arif and Os Keyes. 2022. Vulnerability, Trust and AI. In *Proceedings of Workshop on Trust and Reliance in AI-Human Teams at CHI 2022 (TRAIT)*. 1–7.
[5] Annette Baier. 1986. Trust and Antitrust. *Ethics* 96, 2 (1986), 231–260.
[6] Annette Baier. 1991. "Trust", the Tanner Lectures on Human Values. *Princeton: Princeton University* (1991).
[7] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 167–176.
[8] Veronica Barassi. 2021. The Human Error of Artificial Intelligence. https://www.agendadigitale.eu/cultura-digitale/the-human-error-of-artificial-intelligence/.
[9] The Chartered Institute for IT BCS. 2020. The public don't trust computer algorithms to make decisions about them, survey finds. https://www.bcs.org/articles-opinion-and-research/the-public-dont-trust-computer-algorithms-to-make-decisions-about-them-survey-finds/.
[10] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development* 63, 4/5 (July–Sept. 2019), 4.
[11] Ruha Benjamin. 2014. Race for Cures: Rethinking the Racial Logics of 'Trust'in Biomedicine. *Sociology Compass* 8, 6 (2014), 755–769.
[12] Ruha Benjamin. 2016. Informed Refusal: Toward a Justice-based Bioethics. *Science, Technology, & Human Values* 41, 6 (2016), 967–990.
[13] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity Press.
[14] Sebastian Benthall and Bruce D Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 289–298.
[15] Tithi Bhattacharya. 2017. *Social Reproduction Theory: Remapping Class, Recentering Oppression.* Pluto Press.
[16] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The forgotten margins of AI ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 948–958.
[17] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv preprint arXiv:2004.07213* (2020).
[18] Ankur Chattopadhyay, Abdikadar Ali, and Danielle Thaxton. 2021. Assessing the Alignment of Social Robots with Trustworthy AI Design Guidelines: A Preliminary Research Study. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*. 325–327.
[19] Amy Clair, Jasmine Fledderjohann, and Bran Knowles. 2021. *A Watershed Moment for Social Policy and Human Rights?: Where Next for the UK Post-COVID.* Policy Press.
[20] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.
[21] Kate Crawford. 2021. *Atlas of AI.* Yale University Press.
[22] K Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* 139 (1989), 8.
[23] Angela Y Davis. 2011. *Are Prisons Obsolete?* Seven Stories Press.
[24] Angela Y Davis and Cassandra Shaylor. 2001. Race, Gender, and the Prison Industrial Complex: California and Beyond. *Meridians* 2, 1 (2001), 1–25.
[25] Jason D'Cruz. 2015. Trust, Trustworthiness, and the Moral Consequence of Consistency. *Journal of the American Philosophical Association* 1, 3 (2015), 467–484.
[26] Catherine D'ignazio and Lauren F Klein. 2020. *Data Feminism.* MIT press.
[27] Joseph Donia. 2022. Normative Logics of Algorithmic Accountability. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 598–598.
[28] Jason D'Cruz. 2020. Trust and Distrust. In *The Routledge Handbook of Trust and Philosophy*. Routledge, 41–51.

[29] Jason R. D'Cruz, William Kidder, and Kush R. Varshney. 2022. The Empathy Gap: Why AI Can Forecast Behavior But Cannot Assess Trustworthiness. In *Proceedings of the AAAI Fall Symposium Series Symposium on Thinking Fast and Slow and Other Cognitive Theories in AI.*

[30] Steve Epstein. 2007. *Inclusion: The Politics of Difference in Medical Research.* University of Chicago Press.

[31] Kent Beck et. al. 2001. Manifesto for Agile Software Development. https://agilemanifesto.org/.

[32] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.

[33] Paul Farmer. 2004. *Pathologies of Power: Health, Human Rights, and the New War on the Poor.* Vol. 4. Univ of California Press.

[34] Sina Fazelpour and Maria De-Arteaga. 2022. Diversity in sociotechnical machine learning systems. *Big Data & Society* 9, 1 (2022), 20539517221082027.

[35] Silvia Federici. 2019. Social reproduction theory: History, issues and present challenges. *Radical Philosophy* 2, 4 (2019), 55–57.

[36] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. *Available at SSRN 4020557* (2022).

[37] Centre for Data Ethics and Innovation. 2022. Public Attitudes to Data and AI Tracker: Wave 2. https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-2.

[38] Organisation for Economic Cooperation and Development. 2018. OECD creates expert group to foster trust in artificial intelligence. https://www.oecd.org/innovation/oecd-creates-expert-group-to-foster-trust-in-artificial-intelligence.htm.

[39] Organisation for Economic Cooperation and Development. 2023. OECD Working Party on Artificial Intelligence Governance (AIGO). https://oecd.ai/en/network-of-experts.

[40] Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford University Press.

[41] Herbert J Gans. 1994. Positive Functions of the Undeserving Poor: Uses of the Underclass in America. *Politics & Society* 22, 3 (1994), 269–283.

[42] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward Establishing Trust in Adaptive Agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces.* 227–236.

[43] Ben Green. 2020. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 594–606.

[44] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. [n. d.]. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences.*

[45] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 624–635.

[46] John M Johnson and Andrew Melnikov. 2009. The wisdom of distrust: reflections on Ukrainian society and sociology. In *Studies in Symbolic Interaction.* Emerald Group Publishing Limited.

[47] Frederike Kaltheuner, Abeba Birhane, Inioluwa Deborah Raji, Razvan Amironesei, Emily Denton, Alex Hanna, Hilary Nicole, Andrew Smart, Serena Dokuaa Oduro, James Vincent, et al. 2021. *Fake AI.* Meatspace Press.

[48] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward Situated Interventions for Algorithmic Equity: Lessons from the Field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 45–55.

[49] Helen Kennedy. 2020. Should more public trust in data-driven systems be the goal? https://www.adalovelaceinstitute.org/blog/should-more-public-trust-in-data-driven-systems-be-the-goal/.

[50] Helen Kennedy, Susan Oman, Mark Taylor, Jo Bates, and Robin Steedman. 2020. Public Understanding and Perceptions of Data Practices: A Review of Existing Research. *Sheffield: The University of Sheffield* (2020).

[51] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. 2018. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524* (2018).

[52] Goda Klumbytė, Claude Draude, and Alex S Taylor. 2022. Critical Tools for Machine Learning: Working with Intersectional Critical Concepts in Machine Learning Systems Design. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* 1528–1541.

[53] Bran Knowles, Jason D'Cruz, John T Richards, and Kush R Varshney. 2023. Humble AI. *Communications of the ACM (forthcoming)* (2023).

[54] Bran Knowles and John T Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 262–271.

[55] Bran Knowles, John T Richards, and Frens Kroeger. 2022. The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency. *arXiv preprint arXiv:2208.00681* (2022).

[56] Joshua A Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 758–771.

[57] Abhishek Kumar, Tristan Braud, Sasu Tarkoma, and Pan Hui. 2020. Trustworthy AI in the Age of Pervasive Computing and Big Data. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).* IEEE, 1–6.

[58] Nancy K Lankton and D Harrison McKnight. 2011. What Does it Mean to Trust Facebook? Examining Technology and Interpersonal Trust Beliefs. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems* 42, 2 (2011), 32–54.

[59] John D Lee and Katrina A See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human factors* 46, 1 (2004), 50–80.

[60] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[61] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–14.

[62] Brenda Leong and Evan Selinger. 2019. Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 299–308.

[63] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From Principles to Practices. *Comput. Surveys* 55, 9 (2023), 1–46.

[64] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence* 4, 8 (2022), 669–677.

[65] Q Vera Liao and S Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* 1257–1268.

[66] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* 2103–2113.

[67] Joao Marques-Silva and Alexey Ignatiev. 2022. Delivering Trustworthy AI through formal XAI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12342–12350.

[68] Erich Hatala Matthes et al. 2015. On the Democratic Value of Distrust. *Journal of Ethics and Social Philosophy* 9, 3 (2015), 1–6.

[69] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An Integrative Model of Organizational Trust. *Academy of management review* 20, 3 (1995), 709–734.

[70] Dan McQuillan. 2018. Data Science as Machinic Neoplatonism. *Philosophy & Technology* 31, 2 (2018), 253–272.

[71] Dan McQuillan. 2022. *Resisting AI: An Anti-fascist Approach to Artificial Intelligence.* Policy Press.

[72] José Medina. 2012. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations.* Oxford University Press.

[73] Robert K Merton. 1957. *Social Theory and Social Structure.* Free Press.

[74] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 735–746.

[75] Jakob Mökander and Luciano Floridi. 2021. Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines* 31, 2 (2021), 323–327.

[76] Richa Nagar. 2019. *Hungry Translations: Relearning the World through Radical Vulnerability.* University of Illinois Press.

[77] Safiya Umoja Noble. 2018. *Algorithms of oppression.* New York University Press.

[78] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. 2022. Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* 679–690.

[79] High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[80] Onora O'Neill. 2002. Reith Lectures 2002: A Question of Trust. Lecture 2: Trust and Terror. *BBC Reith Lect* (2002).

[81] Onora O'Neill. 2002. Reith Lectures 2002: A Question of Trust. Lecture 3: Called to Account. *BBC Reith Lect* (2002).

[82] Joon Sung Park, Danielle Bragg, Ece Kamar, and Meredith Ringel Morris. 2021. Designing an Online Infrastructure for Collecting AI Data From People With Disabilities. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 52–63.

[83] Seeta Peña Gangadharan and Jędrzej Niklas. 2019. Decentering technology in discourse on discrimination. *Information, Communication & Society* 22, 7 (2019), 882–899.

[84] Eric Ries. 2011. *The Lean startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.* Crown Business.

[85] Dorothy Roberts. 1997. *Killing the Black Body: Race, Reproduction, and the Meaning of Liberty*. Pantheon Books.
[86] Dorothy Roberts. 2014. Complicating the triangle of race, class and state: The insights of black feminists. *Ethnic and Racial Studies* 37, 10 (2014), 1776–1782.
[87] Loretta Ross and Rickie Solinger. 2017. *Reproductive Justice: An Introduction*. University of California Press.
[88] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 315–328.
[89] Naomi Scheman. 2020. Trust and Trustworthiness. In *The Routledge Handbook of Trust and Philosophy*. Routledge, 28–40.
[90] Nadine Schlicker and Markus Langer. 2021. Towards Warranted Trust: A Model on the Relation Between Actual and Perceived System Trustworthiness. In *Mensch und Computer 2021*. 325–329.
[91] Keng Siau and Weiyu Wang. 2018. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.
[92] Thomas W Simpson. 2012. What is Trust? *Pacific Philosophical Quarterly* 93, 4 (2012), 550–569.
[93] Anubha Singh and Tina Park. 2022. Automating Care: Online Food Delivery Work During the CoVID-19 Crisis in India. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 160–172.
[94] Richa Singh, Mayank Vatsa, and Nalini Ratha. 2021. Trustworthy AI. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*. 449–453.
[95] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxên, Ángeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 667–678.
[96] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 272–283.
[97] Lisa van der Werff, Alison Legood, Finian Buckley, Antoinette Weibel, and David de Cremer. 2019. Trust motivation: The self-regulatory processes underlying trust decisions. *Organizational Psychology Review* 9, 2-3 (2019), 99–123.
[98] Kush R. Varshney. 2022. *Trustworthy Machine Learning*. Independently Published, Chappaqua, NY, USA.
[99] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567.
[100] Anne L Washington and Rachel Kuo. 2020. Whose Side are Ethics Codes On? Power, Responsibility and the Social Good. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 230–240.
[101] Algorithm Watch. 2018. High-Risk Citizens. https://algorithmwatch.org/en/high-risk-citizens/.
[102] Jean Watson. 1997. The Theory of Human Caring: Retrospective and Prospective. *Nursing Science Quarterly* 10, 1 (1997), 49–52.
[103] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.
[104] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* 21 (2019), 89–103.
[105] Roberto V Zicari, John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, et al. 2021. Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society* 2, 2 (2021), 83–97.