Journal of Artificial Intelligence Research 79 (2024) 777-799 Submitted 09/2023; published 03/2024

Performative Ethics From Within the Ivory Tower: How CS Practitioners Uphold Systems of Oppression

Zari McFadden ZMCFADD@NCSU.EDU
Lauren Alvarez LALVARE@NCSU.EDU
North Carolina State University, 890 Oval Dr.
Raleigh, NC 27606 USA

Abstract

This paper analyzes where Artificial Intelligence (AI) ethics research fails and breaks down the dangers of well-intentioned but ultimately performative ethics research. A large majority of AI ethics research is criticized for not providing a comprehensive analysis of how AI is interconnected with sociological systems of oppression and power. Our work contributes to the handful of research that presents intersectional, Western systems of oppression and power as a framework for examining AI ethics work and the complexities of building less harmful technology; directly connecting technology to named systems such as capitalism and classism, colonialism, racism and white supremacy, patriarchy, and ableism. We then explore current AI ethics rhetoric's effect on the AI ethics domain. We conclude by providing an applied example to contextualize intersectional systems of oppression and AI interventions in the US justice system and present actionable steps for AI practitioners to participate in a less performative, critical analysis of AI.

1. Introduction

This article addresses the shortcomings of AI ethics research [1] while presenting a primer for AI practitioners to make critical connections between their work and society. Recent ethics studies have presented perspectives on naming individual systems of oppression (Mohamed, Png, & Isaac, 2020; Munn, 2023), harmful theoretical abstractions (Morley, Elhalal, Garcia, Kinsey, Mökander, & Floridi, 2021; Birhane, Ruane, Laurent, S. Brown, Flowers, Ventresque, & L. Dancy, 2022), and the importance of critical discussion (Waelen, 2022; Schultz & Seele, 2023; van Maanen, 2022; Hampton, 2021). We argue that critical self-reflexivity through a clear understanding of societal systems of oppression and ethics is necessary for more productive research on AI ethics, as supported by (Birhane et al., 2022; Hampton, 2021; Carey & Wu, 2022; Weinberg, 2022), but is often understated by abstracting the pervasive impact of systemic oppression from technology and its creators. We

1. We will use AI ethics and AI fairness as well as researchers and practitioners interchangeably in congruence with current rhetoric. Perhaps the domain language will become more representative of the nuances between ethics and fairness. Still, until then, we refer to AI ethics researchers to include all people involved in developing sociotechnical tools. Including but not limited to those within trustworthy AI, fair AI, interpretable AI, responsible AI, transparent AI, interpretable/explainable AI, AI for social good, and other ethics-related research domains. However, this is not a comprehensive list and is directed at all sociotechnical researchers. We would also like to note the difference between AI ethics and philosophical ethics. We are not theorizing more ethical concepts or interpretations of AI. We are arguing for critical engagement of the scholarly material already available and applying that material to the sociotechnical impacts of technology.

position our argument through an abolitionist framework. We highlight the differences between abolition and reform, where abolition requires creative reimagining of the world as it exists, while reform focuses on making improvements while keeping the world as it is.

Those socialized under Westernized systems of oppression such as patriarchy, white supremacy, colonization, and capitalism are direct products of these systems. A 2022 study (Jakesch, Buçinca, Amershi, & Olteanu, 2022) details the discrepancy between AI practitioners and the general public about the importance of AI, with the general public being more concerned. Along with a higher value for AI ethics, there is also a difference between white and non-white AI ethics practitioners experiencing racial discrimination in the workplace. The 2020 People of Color in Tech Report surveying 1207 professionals in Big Tech (Birch & Bronson, 2022) reported 42.6% of people of color expressed issues due to race/ethnicity compared to 19% of white people (Ton, 2020). This statistic is supported by what Bonilla-Silva describes as 'white habitus' in which *"racialized, uninterrupted socialization process that conditions and creates Whites' racial taste, perceptions, feelings and emotions, and their views on racial matters"* (Bonilla-Silva, 2022). If the non-white engineers who are helping build this technology are experiencing discrimination amongst their peers, what does that say about the discriminatory outcomes of the technology being built? We take a position of abolition, recognizing that those in positions of power desire to maintain that power even if they make small concessions to improve certain outcomes.

The rampant perpetuation of systemic injustice through *algorithmic oppression* by Computer Science (CS) practitioners despite having access to resources and the privilege to learn is a choice, even if unintentional. This choice has violent consequences, particularly for the most marginalized populations (Binns & Kirkham, 2021). With all this knowledge accessible to those in academia and Big Tech, there is a social responsibility to do better, think more critically, and create positive change (Cooper, Moss, Laufer, & Nissenbaum, 2022; Waelen, 2022; Munn, 2023; Mittelstadt, 2019). Computing and theoretical ethics allow researchers to abstract and "discover" discrimination in models. This theoretical abstraction removes accountability from the researchers and places it on "black-box" models. Unsurprisingly, simple AI models have observed obvious discriminatory relationships and are repeatedly presented in books (Benjamin, 2020; Noble, 2018; Eubanks, 2018; O'neil, 2017) and landmark cases (Buolamwini & Gebru, 2018; Angwin, Larson, Kirchner, & Mattu, 2016; Obermeyer, Powers, Vogeli, & Mullainathan, 2019; Keyes, 2018).

This article begins with key definitions and explicitly connects the societal systems of oppression and power to AI fairness. **This is not an all-inclusive paper, but rather a primer to serve as a "crash-course" resource to frame future critical AI ethics research.** We then discuss the current AI ethics rhetoric and present a contextualized case study example. The following section expands on our abolitionist position and comments on current AI regulations' shortcomings. The final section concludes by discussing less performative, more critical AI ethics research, and our suggestions for future work.

2. AI Ethics' Position Within Systems of Oppression & Power

In this section, we explicitly define and denote ten systems of oppression and power motivating this work, what they look like in practice, and their implications on both a local and global scale. The motivating systems of oppression and power include, but are not

limited to, *capitalism and classism; colonialism, colonization, and imperialism; racism and white supremacy; sexism and patriarchy; and ableism* defined in Table 1 and Table 2. Yes, this seems like a lot of -isms, but each of these -isms interact, shaping how each person experiences and interacts with the world. We acknowledge these are interlocking systems of power that have differing consequences for different intersections of identity (Crenshaw, 1989; Collins & Bilge, 2020), but for clarity purposes, we save that for further discussion in future work. To understand the impact of intersectional systems of oppression, one must begin with understanding singular systems first. Given the lack of papers within the AI ethics domain discussing systems of oppression at all, referring to each system individually is the current scope of this work. Understanding these systems sets the base for critically analyzing technology and AI ethics. The next step in understanding is recognizing that a truly equitable society would require abolishing these systems of oppression because as long as they exist, people will continue to be oppressed. The technology we build will just uphold and exacerbate that oppression. This section defines the aforementioned systems, why they are relevant, and how they connect to technological development more broadly. In addition, this section will parse and add nuance to common rhetoric related to AI ethics and building socio-technical tools.

2.1 Capitalism and Classism

We argue **capitalism** is the first and most important word to analyze and dissect because it is the most fundamental system shaping our lives, and the following systems, to be discussed, are utilized towards the goals of capitalism. Mandel summarizes Karl Marx's definition of capitalism as *"the ruthless and irresistible impulse to growth which characterizes production for private profit and the predominant use of profit for capital accumulation"* (Marx, 1990). Capitalism, as a system of power that directly or indirectly impacts the global society (Lechner & Wallerstein, 2015; Morley, 1989), is more complicated than space allows us to explore here. However, it is crucial to understand the implications of how the United States' economic system, the locus of most big tech companies, is structured to value profit over everything else. When profit is valued over people, the primary concerns are those that increase profit; this extends into technology and debates about ethics. For example, Access Now, an international digital and human rights organization, resigned in protest from Partnership on AI (PAI), a consortium of Big Tech companies including Apple, Amazon, Facebook, Google, IBM, and Microsoft (Johnson, 2020). As Access Now stated in a letter about their resignation, they *"did not find that PAI influenced or changed the attitude of member companies or encouraged them to respond to or consult with civil society on a systematic basis"* (Johnson, 2020). Under capitalism, capitalists and corporations are primarily accountable to the stakeholders that rely on them for profit, with limited accountability to government regulation.

As Césaire notes, *"capitalist society, at its present stage, is incapable of establishing a concept of the rights of all men, just as it has proved incapable of establishing a system of individual ethics"* (Césaire, 1972). Any conversation concerned with the impact of AI and technology on society must also be concerned with the impacts of capitalism and the reality of economic profit over consumer impact. Capitalism is driven by profit and the accumu-

Table 1: Systems of Oppression and Power Definitions.

Term	**Definition**
Capitalism	"an economic system based on the private ownership of the means of production. Capitalism is typically characterized by extreme distributions of wealth and large differences between the rich and the poor." (Hill Collins, 2000) "the capitalist mode of production, the seizure of the means of production by capital, which has become predominant in the sphere of production" (Marx, 1990)
Classism	"negative attitudes, beliefs, and behaviors directed toward those with less power, who are socially devalued" (Lott, 2012)
White Supremacy	"an ideology that presents the ideas and experiences of Whites as normal, normative, and ideal" (Hill Collins, 2000)
Racism	"a system of unequal power and privilege where humans are divided into groups or "races" with social rewards unevenly distributed to groups based on their racial classification. Variations of racism include institutionalized racism, scientific racism, and everyday racism. In the United States, racial segregation constitutes a fundamental principle of how racism is organized." (Hill Collins, 2000)
Patriarchy	"a political-social system that insists that men are inherently dominant, superior to everything and everyone deemed weak, especially females, and endowed with the right to dominate and rule over the weak and to maintain that dominance through various forms of psychological terrorism and violence" (hooks, 2004)

Table 2: Systems of Oppression and Power Definitions Cont.

Term	**Definition**
Sexism	when one is "discriminated against on the basis of sex" (hooks, 1984) "sexism should be understood primarily as the "justificatory" branch of a patriarchal order, which consists in ideology that has the overall function of rationalizing and justifying patriarchal social relations." (Manne, 2017)
Ableism	"a set of beliefs or practices that devalue and discriminate against people with physical, intellectual, or psychiatric disabilities and often rests on the assumption that [people with disabilities] need to be 'fixed' in one form or the other" (Smith, 2023)
Colonialism	"a form of domination - the control by individuals or groups over the territory and/or behavior of other individuals or groups" (Horvath, 1972)
Colonization	can be thought of as three different types (Horvath, 1972): 1. "colonization in which the dominant relationship between the colonizers and the colonized is extermination of the latter" 2. "colonization in which assimilation is the relationship between the colonizers and the colonized" 3. "colonization in which settlers neither exterminate nor assimilate the indigenes"
Imperialism	"a form of intergroup domination wherein few, if any, permanent settlers from the imperial homeland migrate to the colony" (Horvath, 1972)

lation of capital [2]. What is prioritized for capitalists is often not suitable for the working class (or proletariat). This idea is captured by **classism**. For example, producing items as cheaply and quickly as possible is in the best interests of the capitalist class because of the increase in profit, but it is not suitable for the laborers who are paid less than living wages and are forced to work in dangerous conditions (Epatko, 2018; Lee, 2022; Perraudin, 2019).

2. "value (initially in the form of money) becoming an independent operator in the pores of a non-capitalist mode of production" (Marx, 1990)

"value constantly increased by surplus-value, which is produced by productive labour and appropriated by capitalists through the appropriation of the commodities produced by the workers in factories owned by capitalists"

Constant production and overconsumption also have a drastic environmental impact that continues as climate change progresses (Panagiotopoulou & Chryssolouris, 2022). Technology continues to exacerbate the environmental (Nair, 2023; Report, 2023) and human consequences (Perrigo, 2023; Dzieza, 2023) involved with training large language models, sourcing materials to build technology, and planned obsolescence that forces consumers to constantly buy new technology.

2.2 Colonialism, Colonization, and Imperialism

Most AI ethics work and the societal impact of technology has been centered around the Western perspective; this is a result of **colonialism** and ignores the treachery that remains an invisible undercurrent in the building blocks of our technology. The United States' foreign policy and the history of Europe's direct and indirect **colonization** and **imperialism** are out of the scope of this paper. Still, their historical legacy is the basis of the technological reality we currently live in. There would be no computers or phones without the exploitation of the Global South where children as young as six in the Democratic Republic of the Congo (DRC) are risking their lives to mine cobalt (Kara, 2018), providing the very minerals that are necessary for computer production. Furthermore, the present day would not exist as it does if it were not for Leopold II establishing the DRC as a colony of Belgium in the 19th century, leading to the deaths of over 10 million people (Hochschild, 2022). Colonialism (and the requirement of profit by capitalism) is the reason why there exists a Global North and a Global South. As Morley notes, *"the U.S. state, as an imperial state shaped and controlled by outward-looking capital, assumes a multiplicity of tasks to facilitate the goals of its outwardly oriented capital class"* (Morley, 1989). The connections between capitalism, colonialism, and imperialism can be seen in a company like OpenAI outsourcing their labor to Kenyan workers for only $2 an hour to make ChatGPT "less toxic" (Perrigo, 2023). A company with plans to raise funds to reach a $29 billion valuation is exploiting the labor of Black people from an under-resourced country to fix its violent, racist, and sexist technology for $2 an hour without acknowledgment of the lasting psychological harms. That is one of the clearest, contemporary examples of how each of these interlocking systems of oppression collide in the tech space.

2.3 Racism and White Supremacy

Racism works hand-in-hand with white supremacy to uphold a social order that privileges whiteness over everything else. Although race is a social construct with no biological basis (Smedley & Smedley, 2005), it has tangible impacts on the lives of white and non-white people. For white people, whiteness is the hegemonic [3] norm, an invisible background force that allows for the maintenance of privilege to the exclusion of everyone else. Whiteness as a political construct was used to justify colonization and slavery (Robinson, Sojoyner, & Willoughby-Herard, 1983). People like Charles Darwin and previous world leading statistician R.A. Fisher (Pearl & Mackenzie, 2018), used race as an underpinning of evolutionary theory to codify supremacy of white people over non-white people (phrenology): *"Although Darwin's book was criticized for its stance against the church, the British empire used it to*

3. "the social, cultural, ideological, or economic influence exerted by a dominant group" (Merriam-Webster, 2023)

justify colonialism by claiming that those subjected under its rule were scientifically inferior and unfit to rule themselves, with British anthropologists like James Hunt using Darwin's theory to justify slavery in papers such as The Negro's Place in Nature (1863)" (Gebru, 2019). The ruling political context (regardless of individual personal belief) of **white supremacy** is a socially dominant force fundamental to the prominent scientific discoveries, extending to AI and other technologies. Hanna et al. note *"despite the risk of reifying the socially constructed idea called race, race does exist in the world, as a way of mental sorting, as a discourse which is adopted, as a social thing which has both structural and ideological components. In other words, although race is social [sic] constructed, race still has power"* (Hanna, Denton, Smart, & Smith-Loud, 2020). There have been many examples of various technologies performing best on white people and worst on non-White, but primarily Black and Latine people. Some examples include facial recognition (Raji, Gebru, Mitchell, Buolamwini, Lee, & Denton, 2020), recidivism prediction (Angwin et al., 2016), and healthcare (Obermeyer et al., 2019; Grant, 2022). Ignoring race in AI ethics research and development areas is a blatant upholding of racism and the hegemonic norm of white supremacy. Furthermore, performatively naming racism or intersectionality in AI ethics research does not automatically denote critical engagement. Racism, white supremacy, intersectionality, etc. should not just be treated as buzzwords, but instead as markers of structural oppression with tangible consequences.

2.4 Patriarchy and Sexism

Patriarchy is a structural system that is upheld by and harmful to men, women, children, and non-binary people alike. Violence is commonly defined as *"behavior involving physical force intended to hurt, damage, or kill someone or something"* (Google, 2023), but reducing violence to just physical force does a disservice to the other forms that violence takes (Bufacchi, 2005); poverty, police states, microaggressions, verbal and emotional abuse, etc. are all examples of non-physical violence. Sexism is a consequences of patriarchy with bell hooks noting that *"under capitalism, patriarchy is structured so that sexism restricts women's behavior in some realms even as freedom from limitations is allowed in other spheres"* (hooks, 1984). A concern consistently at the forefront of the tech space is getting more women involved in CS (Aguilar, 2022; Oi, 2022). What is not often included in these calls is the explicit naming of patriarchy and patriarchal violence as one of the mechanisms impacting the experiences women and girls have in male-dominated tech spaces, not including other dynamics of race, class, or ability.

Patriarchy as a framework is genderless (on Gender-Based Violence, 2018; hooks, 2004), though it more directly benefits men and more directly harms women and non-binary people. It is also present in the type of technical work deemed worthy of respect. In a patriarchal society, things coded as feminine or designated to women's space are regarded as less valuable regardless of the gender of the person participating. Terrence Real defines this in terms of his role as a therapist, but the same idea can be extrapolated to other areas coded as traditionally feminine: *"all therapists, under patriarchal mores, are coded as female, and as such they are subject to the same devaluing and intimidation as are traditional wives"* (Real, 2003). Direct examples of sexist AI include showing gender-biased ads (Benjamin, 2020), the fetishization of Black and Latine women in search engines (Noble, 2018), women not

being considered in research (Perez, 2019; Harding, 1991), the reinforcement of gendered stereotypes through virtual assistants (Loideain & Adams, 2020; Chin & Robison, 2020; West, Kraut, & Ei Chew, 2019), and misgendering/outing non-binary people (Costanza-Chock, 2018; Keyes, 2018).

2.5 Ableism

Ableism is a system that operates with an air of invisibility, especially for those who are able-bodied, and often leads to the death of people with disabilities who are thought of as disposable to society. Williams et al. asked student participants about the ethics of various hypothetical medical situations and highlighted ableism as incredibly prevalent:

> *"Though we found most participants recognized race, gender, and class biases as threats to the ethics of the scenarios they evaluated, only one participant explicitly warned the eugenic potential of these systems, and no participants clearly identified disabled people as a social class vulnerable to unique mechanisms of discrimination. This gap in student understanding is further illustrated by the ways that they implicitly identify "society" as medical practitioners, institutions, families, and friends—those social agents whom [sic] traditionally have substantial power over disabled people's access to care, culture, and life."*(Williams, Smarr, Prioleau, & Gilbert, 2022).

The scenarios positioned as hypotheticals in the study were actual events such as the use of predictive algorithms to determine the allocation of resources (Pourhomayoun & Shakibi, 2021; Shanbehzadeh, Yazdani, Shafiee, & Kazemi-Arpanahi, 2022). This technology is especially dangerous for people with disabilities because their disabilities are used to put them in higher risk categories that are often used as support for withholding care to the benefit of abled-bodied people. This was especially seen in the way the media addressed COVID-19 mortality rates: *"Public health officials, journalists, and politicians of pretty much every variety have said explicitly, or implied, that whatever current form of covid is under discussion can be regarded as at least a little less worrying because it mainly sickens and kills elderly, chronically ill, and disabled people"* (Pulrang, 2022). When such explicit language is used in the media in reference to the lives of people with disabilities, who can fully know the impact of inscrutable algorithms being used to determine the allocation of care? Who is being sacrificed and relegated as less valuable? Other examples of ableism in AI are present in papers about participatory fixes excluding people with disabilities (Sloane, Moss, Awomolo, & Forlano, 2020), the usage of sensitive medical data for people with disabilities (Binns & Kirkham, 2021), and other critique papers (Shew, 2020; Tilmes, 2022).

3. A More Expansive Understanding of Bias, Fairness, and Ethics

The next crucial point to acknowledge is the impact of context and rhetorical nuances on research development. Words like bias, fairness, and ethics are nearly ubiquitous in discussions about the impact of various technologies, but the ways each of those words is used in the tech space is often different from their traditional meanings. Fairness and fairness interventions have been used as an operationalization tool for ethics. However,

we argue that fairness is inaccurately operationalizing ethics. This section explores the aforementioned words, how their meanings change depending on the context, and what that means for us as engineers and technologists.

3.1 The Misuse of Bias and Fairness

Fairness and bias are common words used when considering technology's impact and repairing its negative lasting effects (Ehsan, Singh, Metcalf, & Riedl, 2022a). Other researchers have commented on the various definitions of fairness, such as Hampton's reference to over 21 definitions presented in (Tal, Batsuren, Bogina, Giunchiglia, Hartman, Loizou, Kuflik, & Otterbacher, 2019a). In both senses of the word, bias has to do with a lack of fairness, meaning that fairness and bias are essentially two sides of the same coin. The problem with these words' ubiquitous nature is their ambiguity and superficial relation to the real-world impact they abstractly represent. For there to be a lack of equality, there has to be a comparable original amount. What lens is the world being viewed through to determine what is actually fair, and to whom does this fairness apply? Have there ever been moments throughout human history of total fairness? If not, how do we know what to strive for? What does fairness mean in an unfair world? As Berhane et al. note, *"the treatment of concepts such as fairness and bias in abstract terms is frequently linked to the notion of neutrality and objectivity, the idea that there exists a 'purely objective' dataset or a 'neutral' representation of people, groups, and the social world independent of the observer/modeller"* (Birhane et al., 2022). With white, heterosexual, male, able-bodied perspective being socially constructed as the hegemonic norm, the spectrum of how fairness is quantified should be heavily reviewed. We point to this paper for further discussion on specific fairness metrics (Carey & Wu, 2022).

As mentioned by multiple scholars (Buyl & De Bie, 2022; Tal, Batsuren, Bogina, Giunchiglia, Hartman, Loizou, Kuflik, & Otterbacher, 2019b), AI ethics work has no widely accepted definition of fairness. This could be defended by needing more context or understanding of the application for which fairness is being assessed. However, the lack of a widely accepted definition and contradiction of current methods creates too much nuance that can have unanticipated negative effects such as Fairness Gerrymandering, or the Simpson's Paradox (Kearns, Neel, Roth, & Wu, 2018). A lack of a clear fairness definition can lead to explicitly harmful consequences because the current state of success would be "fair enough", "fairer than previously" or as "fair as it could be", which is not a concrete or acceptable answer (Rakova, Yang, Cramer, & Chowdhury, 2021). AI and machine learning aim for accurate generalizations and, ideally, a concept of *group fairness* that considers *individual fairness* (Berk, Heidari, Jabbari, Joseph, Kearns, Morgenstern, Neel, & Roth, 2017; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021) which can, but not always, connect to distributive justice (Gabriel, 2022). However, with so much confusion around the definition of fairness, fairness interventions do not operationalize a specific AI ethics subfield.

3.2 Glorification & Abstraction of Ethics

Ethics is a popular term in the AI space that acts as a catch-all term for considering the morality and impact of technology. However, this perspective oversimplifies the philosoph-

ical field of ethics and glorifies the field as a designated source of truth about what is right rather than a theoretical abstraction. Ethical theorizing allows AI practitioners to abstract ethics into a speculative exercise rather than a practical problem with real, violent impacts (Birhane et al., 2022). Our goal is to shift focus from abstraction to structural connections, and discuss the dangers of performative ethics.

Ethics as a philosophical field is composed of three different categories: normative ethics, applied ethics, and metaethics (of Wisconsin Madison, 2023). AI ethics tends to fall in the category of metaethics which *"investigates where our moral values, language, and principles come from and what they mean; it is concerned with 'what is morality?' rather than 'what is moral?"'* (of Wisconsin Madison, 2023). This distinction is important because addressing the ethics of technology without specifying the ethics framing collapses together different aspects of ethical work. Furthermore, as Siapka notes, *"Although much of the literature focuses on identifying a suitable ethical framework for the development and deployment of AI, there have not been comparable efforts for the identification or construction of frameworks within which to carry out a second-order reflection on AI ethics"* (Siapka, 2022). AI metaethics invites us to think about whether there is such a thing as fairness, who defines fairness, and whether fairness can be a realistic goal. When it comes to thinking about the ethics of various technologies, researchers must work to define the various frameworks that will be used to classify and analyze the ethical impacts of technology. As Birhane et al. note,

> *"canonical Western approaches to ethics, from deontology to consequentialism, at their core strive for such universal and generalizable theories and principles 'uncontaminated' by a particular culture, history, or context. Underlying this aspiration is the assumption that theories and principles of ethics can actually be disentangled from contingencies and abstracted in some form devoid of context, time, and space"* (Birhane et al., 2022).

When glorifying ethics as a field and looking to it as a guiding light for answers about morality, there has to be intentionality about critically understanding the underlying philosophical perspectives. Using ethics that are rooted in the same systems of power mentioned above is not enough to determine rightness *"given the ways that existing AI ethics literature builds on the circulation of existing philosophical inquiry into ethics, the reproduction of the exclusion of marginalized philosophies and systems of ethics in AI ethics is unsurprising, as is the maintenance of the white, Western ontologies upon which they are based"* (Birhane et al., 2022). Suppose there is a desire to generate specific fields dedicated to AI ethics. Those fields must be intentionally created to broaden the scope of ethics (and the subfields that compose it) beyond the Western context. We point to these works for examples of consideration of non-Western perspectives (Hongladarom & Bandasak, 2023; Abebe, Aruleba, Birhane, Kingsley, Obaido, Remy, & Sadagopan, 2021).

4. Liberation Cannot Be Operationalized Under Systems of Oppression

Historically, computing has been able to abstract itself from its embedded social construction and only recently has been critiqued for the resulting harms. To deconstruct systems of oppression that are reified through algorithmic oppression, we argue that socio-technical

researchers must have a common foundational base of sociology, ethics, and self-reflexivity. A prevalent limitation in AI ethics work is the avoidance, or inability, to explicitly state the critical structures that shape the world and examine the impacts of sociotechnical systems on society (Ehsan et al., 2022a). Without systemic analysis, we claim work dedicated to positively improving the impact of technology on society will be performative at best and reify systems of oppression at worst. The following sections demonstrate the difference between reform and abolition, and the importance of contextualizing research with systems of oppression.

4.1 Abolitionist Framework

Abolitionist movements can be traced throughout history to the goal of abolishing slavery. In more recent times, calls for abolition have moved beyond abolishing chattel slavery to calling for the abolishing of police and prisons. Notably, work from Angela Davis (Davis, 2003), Ruth Wilson Gilmore (Gilmore, 2022), and others challenge us to question the necessity of cages and punitive responses to societal failure. Abolition requires creativity and imagination. As Davis notes regarding prison abolition, *"it has become so much a part of our lives that it requires a great feat of the imagination to envision life beyond the prison..."* (Davis, 2003). Abolition requires the conceptualization of a world outside of current systems of oppression.

While there is an increase in sociotechnical acknowledgments and inclusive vocabulary (Birhane et al., 2022; Munn, 2023; Mittelstadt, 2019), much of the existing AI ethics research makes suggestions of reform rather than abolition. For example, a paper about decolonizing AI (Mohamed et al., 2020) acknowledges the *"coloniality of power"* and argues for structural decolonization *"that seeks to undo colonial mechanisms of power, economics, language, culture, and thinking that shapes contemporary life,"* but does not make explicit calls for abolition. Instead, the authors suggest three modes of *"reciprocal tutelege"*, which are dialogue, documentation, and design. These suggestions can be seen as ideals of reform and the initial stages of Harro's Cycle of Liberation (Harro, 2000). Reform solutions fall in line with the ideals of techno-solutionism, where technology is seen as the most optimal solution for all problems as long as the right problem set is outlined and the technologies exist (Morozov, 2013; Gardner & Warren, 2019). Although improvements can be made to the outcomes of AI systems through reform approaches, focusing on reform alone is not enough to address systems of oppression and can often lead to the further reification of power and oppression.

Reform ideals operate from a position that if we improve certain aspects of the system, then the problems of oppression and discrimination will be fixed. Rodriguez notes that *"to reform a system is to adjust isolated aspects of its operation in order to protect that system from total collapse, whether by internal or external forces"* (Rodriguez, 2020). In the AI ethics space, this looks like the implementation of datasheets for datasets, focuses on AI governance and policy, community-centered design, etc. While each of these contributions are important and help make the technology we build a little better, a little more fair, we argue to extend beyond reforms. The goal of abolition is a lofty one and it feels like too big, too unreasonable of a solution, but so did abolishing slavery. As Gilmore conceives, abolition is not about having all the answers of how,it is about magic: *"meaning we don't*

yet know how, which is what magic is, what we don't know how to explain yet" (Gilmore, 2022). Abolition requires constant reimagining of the world as it exists and acknowledgment that we cannot band-aid solution or *"undo"* our way out of structural harm.

4.2 Abolition, Colorblind Racism, and the Role of Engineers

In his book *Racism Without Racists: Colorblind Racism and the Persistence of Racial Inequality in the United States*, Eduardo Bonilla-Silva explores the maintenance of racial hierarchy and inequality that continues despite the lessened experience of direct racism. He clarifies *"I have argued elsewhere that contemporary racial inequality is reproduced through 'New Racism' practices that are subtle, institutional, and apparently non-racial. In contrast to the Jim Crow era, where racial inequality was enforced through overt means (e.g., signs saying 'No Niggers Welcomed Here' or shotgun diplomacy at the voting booth), today racial practices operate in a 'now you see it, now you don't fashion"* (Bonilla-Silva, 2022). With technology, colorblind racism [4] (Carr, 1997) becomes even more obscure and expansive through machines built to be "black boxes" with often little explanation or accountability for the decisions made (Pedreschi, Giannotti, Guidotti, Monreale, Ruggieri, & Turini, 2019). The idea of "colorblindness" can be extended outside of race (i.e., ableism, classism) to include the often invisible privileges of power structures that systematically disenfranchise marginalized groups. Thus, CS practitioners are often operating from and participating in two layers of invisibility: (1) the invisibility of the ways they benefit from various systems of oppression and are invisibly passing those beneficial assumptions to the technology they build and (2) the invisibility of abstracting ethics and redirecting accountability to a "discriminatory machine" (i.e., the black box). As Hanna et al. note, *"treating race [or any marginalized aspect of identity] as an attribute, rather than a structural, institutional, and relational phenomenon, in turn, serves to minimize the structural aspects of algorithmic unfairness"* (Hanna et al., 2020). Without a critical understanding of systems of oppression, high-risk AI is abstracted away from the real violence it can and does cause, and the responsibility falls on the creator who ignorantly [5] pursued a solution without gathering the necessary socio-technical background to fully understand the problem.

In their work on coliberative consciousness in CS pedagogy, Williams et al. show that despite having an understanding of social justice issues, many students still envision ethical issues with AI systems as a result of *"biased datasets and human mis/trust factors, rather than as problems of design and purpose"* supporting the idea of a kind of colorblindness present among engineers and technologists (Williams et al., 2022). They explain, *"Even when ethics are directly addressed in CS curriculum, instruction tends to focus around professional, corporate, or legalistic frames of ethical behavior rather than the ethics of developing and deploying systems within their sociotechnical context"* (Williams et al., 2022). There is a present belief that objectivity, merit, or following suggested guidelines within these frames, are the criteria for 'fair' systems. However, as Gebru notes, *"an analysis of scientific thinking in the 19th century, and major technological advances such as automo-*

4. "in terms of racial colorblindness, a person is also choosing to not just see race or skin color, but also the racial disparities, inequities, history of violence and current trauma perpetuated within a racist society" (Library, 2023)
5. We would like to make further connections to *weaponized ignorance*, but for the benefit of conciseness see (McKeever, 2022; Froehlich, 2017).

biles, medical practices, and other disciplines shows how the lack of representation among those who have the power to build this technology has resulted in a power imbalance in the world and in technology whose intended or unintended negative consequences harm those who are not represented in its production" (Gebru, 2019). Without an understanding of the sociotechnical context, which requires understanding societal systems of oppression, many of the AI issues will be limited to how a system is implemented rather than the society in which the implementation is taking place.

We must be clear there is a difference between decreasing a system's measurable disparate impact [6] and whether the system's implementation is working to dismantle or perpetuate the systems of oppression these automated systems operate within. With an abolitionist perspective, there is a better foundational basis to ask questions about the goals of a particular technology, whether that technology should exist, and how the technology interacts with the rest of the world. It also leaves space for realizing that no technology will be enough to fix systems that have been in place for hundreds, if not thousands of years, ridding all of us of the delusion that AI (or any other technological system) will be some kind of savior from all the problems in the world (Birhane et al., 2022; Ahmed, 2004). Addressing these systems takes more than a pithy checklist, an anti-racism course, or diversity training; it requires starting with a much more systemic analysis and understanding of the world and the historical depth of the complexities in which we live.

5. AI in the U.S. Judicial System

In this section, we examine a high-risk domain, discuss the ways in which current research suggestions are shortsighted, and provide an example of how we argue the domain should be contextualized in future work. First, we present the necessary background. A ProPublica article (Angwin et al., 2016) published in 2016 found that the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) used to predict recidivism was racially biased/racist against Black people, predicting that Black people have a higher risk of committing crimes or reoffending than white people. This type of predictive technology has a tremendous capacity for harm because judges can use it to determine sentences, meaning that algorithms influence how long people stay in cages. A group at the Northpointe Inc. research department did a counterstudy to the ProPublica claims (i.e., asserting that the COMPAS risk scales were not racially biased). It illustrated that ProPublica made several statistical and technical errors to reach their conclusions (Dieterich, Mendoza, & Brennan, 2016). Their argument supported by Rudin et al. (Rudin, Wang, & Coker, 2020) is that there is no racial bias because if they remove race as a feature, the results are the same; they do not deny the algorithm has disproportionately negative outcomes for Black people compared to white people (which is what the ProPublica study argues). The ProPublica counterargument is presented to showcase how complicated fairness can be and that statistical fairness metrics do not necessarily represent fairer outcomes.

The larger research discussion fails to critically analyze the dataset and the pervasive impact of automation in this field. The abstraction of mitigating recidivism is digestible for current practitioners, but contextualizing their model is the blind spot (Hagendorff,

6. "when the decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups" (Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017)

2022) and results in reifying systemic, historical oppression through *algorithmic oppression*. Current perspectives on recidivism prediction include improving fairness through dataset manipulation or increasing interpretability, while other perspectives believe that recidivism prediction outcomes are not unfair because the model is statistically accurate based on the data. Both directions ignore the historical, systemic oppression that impacted the people represented in the dataset, and are incredibly ignorant. Focusing on the data abstraction and improving statistical models, removes the larger socio-historical context that models are being integrated into. The result is explicitly discriminatory models that impact real people, and the harms that have passed can never be undone. The algorithmic imprint (Ehsan, Singh, Metcalf, & Riedl, 2022b) will be longstanding, and until a critical, ethical intervention is presented, these algorithms should be removed. There are very few AI researchers (Hampton, 2021) who advocate for the removal of harmful algorithms. From the abolitionist perspective, recidivism automation should be abolished because these systems perpetuate a long history of violence.

5.1 Contextualized Case Study

To further demonstrate our position for critical ethical engagement, we provide a contextualized formula that investigates (1) a rarely acknowledged system of oppression, (2) current research interventions, and (3) the unseen negative impacts of performative AI ethics research without critical analysis. Our applied example examines (1) mass incarceration [7], (2) discriminatory risk models, and (3) automated "justice." While these concepts may not seem to relate directly, we seek to spotlight systems of oppression and how these systems manifest in AI ethics research and related fields, and the dangers of inappropriate, abstracted solutions.

We note that mass incarceration as a system of oppression is rarely stated in AI ethics publications, especially within recidivism prediction research. According to The Sentencing Project, the United States is the world's leader in mass incarceration (Nazgol Ghandnoosh & Nellis, 2022). There are currently *"1,566 state prisons, 102 federal prisons, 2,850 local jails, 1,510 juvenile correctional facilities, 186 immigration detention facilities, and 82 Indian country jails"* (Wagner & Peter, 2022), totaling 6,296 places to hold humans in cages in the U.S. compared to 4,360 higher education institutions (Bouchrika, 2022; Duvernay, 2016), and many private prisons have capitalized on incarcerating Black and Brown men and boys. While the U.S. only makes up 4% of the world's population, it is home to 16% of the world's prisoners (Vera, 2023). Of the people incarcerated, Black, Latine, and Native American people are overrepresented (with Black people most severely overrepresented), while white people are underrepresented. Black people make up only 13% of the U.S. population, but 38% of the incarcerated population, Latine people make up 19% of the U.S. population, but 21% of the incarcerated population, and Native American people make up 0.9% of the U.S. population, but 2% of the incarcerated population, while white people make up 60% of the U.S. population, but only 38% of the incarcerated population. This is despite the fact that crime rates are consistent across communities, with most crimes

7. We highlight capitalism, white supremacy, and colonization are interconnected to mass incarceration and criminalization, and without recognizing the impacts of these systems, the automated systems embedded in these contexts will be harmful. Although we stated that intersectionality was out of scope of this paper, there are certain systems that cannot be removed because systems of oppression are inherently connected.

occurring intra-racially at similar rates (Caldera, 2020) and drug usage being similar across racial backgrounds (Hinton, Henderson, & Reed, 2018; Project, 2018). The history of U.S. policing stemming from the "Slave Patrol" in the early 1700s Carolinas (NAACP, 2021) and the U.S. 13th amendment being a loophole to allow slavery as a punishment for crime is out of scope of this paper, but the issue of racism, white supremacy, and discrimination has been a fundamental aspect of the creation of policing in the United States.

With racialized policing practice in mind, how can we train predictive policing models in fair and unbiased ways when the data used for training is steeped in the history of racist policing? Even if the process by which AI models were being built were completely free of bias, that would not change the reality of policing, and the data it produces, being purposely and historically discriminatory against Black and Latine people. AI, no matter how statistically fair the results are, cannot change centuries of historical precedence and discrimination that continue to this day. Consider two of the interconnecting systems of oppression (mass incarceration and criminalization) impacting racial demographic representation in the industrial prison system, and let's connect these systems to current research suggestions and the violent impact of automating a system without the proper socio-technical background. We mentioned the landmark COMPAS recidivism model, which is one of many recidivism models active in the United States. In fact, 60% of the U.S. population lives in a jurisdiction that actively uses risk assessment tools (Injustice, 2023). By coupling AI to tangible systems of oppression in the United States, the only result is automated systemic injustice and algorithmic oppression. The suggested interventions do not situate the models, nor do they consider the systems of oppression that influence recidivism prediction.

The abolitionist perspective as a starting position asserts that certain inalienable human rights belong to every human being across the globe and that human rights trump the rights of corporations. From there, work can begin to think about who has access to those rights and who does not and how we can build a society (and technology along with it) to grant everyone access to those rights. Then, solutions can be engineered from positions of justice, equality, and equity, recognizing that there is no technical, mathematical solution without dismantling systems of oppression.

6. Discussion

This paper presents a primer for AI practitioners to make critical connections between their work and systems of oppression. This work contributes to related literature regarding critical engagement and applying said material to the sociotechnical impacts of technology. We argue that critical self-reflexivity through a clear understanding of societal systems of oppression and ethics is necessary for more productive AI ethics research, but is often understated through theoretical abstraction. To deconstruct systems of oppression being reified through algorithmic oppression, socio-technical researchers must have a common foundational base of sociology, ethics, and self-reflexivity. As a case study, we choose to examine a high-risk domain, discuss the ways that current research suggestions are shortsighted, and provide an example of how we argue the domain should be contextualized. Our goal is to unmask performative ethics and shift focus from reform to systemic engagement.

A novel contribution of this work is interrogating the difference between reform and abolition in AI ethics research. Much of the current research provides solutions and sug-

gestions that fit within the realm of reform, but we argue for an abolitionist framework. While the reform suggestions that existing research provides are valuable, it is necessary to be aware of how reform can act to deradicalize radical traditions to avoid shaking the boat of power and respectability too much. Crenshaw notes a history of racial liberalism being used to deradicalize racial reform (Crenshaw, 2017) and similar concepts can be extended to the ideas of "ethics washing" (Schultz & Seele, 2023; van Maanen, 2022) by Big Tech corporations. We are pushing for AI ethics research to move in a direction beyond reform alone. As Gilmore states *"the abolition I speak of somehow, perhaps magically, resists division from class struggle and also refuses all the other kinds of power difference combinations that, when fatally coupled, spark new drives for abolition. Abolition is a totality and it is ontological. It is the context and content of struggle, the site where culture recouples with the political; but it is not struggles' form"* (Gilmore, 2022). An abolitionist framework asks for creativity and imagination and struggle. We urge the research community to struggle together to imagine a new world without oppressive systems and the technology that upholds them.

Our work intentionally does not offer concrete solutions. Part of the abolitionist framework is not knowing how to achieve the end goals *right now* and we are okay with this position for now. Our goal for this paper is to participate in the first two steps of cycle of Harro's Cycle of Liberation which is "waking up" and "getting ready" (Harro, 2000). We are waking up CS practitioners to the various systems of oppression and encouraging them to get ready to critically engage with the systems in their research. Like Gilmore, we are looking forward to the magic of imagining new worlds without these systems and figuring out AI/technology's role in that world. We encourage readers to dream and imagine even if there are no actionable solutions just yet. We encourage sitting in the discomfort of not knowing what is next. We encourage grappling with ideas of reform versus abolition and how to make progress without further entrenching the existing systems.

6.1 Future Work

This paper points back at those present in academia and Big Tech because when well-intentioned interventions aren't successful and are consistently reimagined with no significant impact, when does performative ethics research become harmful? "Well-intentioned" isn't enough. Academia is coined the "ivory tower" meaning *"it retains the notion of an authorized body of knowledge, relies primarily on classroom-based (or laboratory) learning, does not stress subjective experience as a legitimate source of knowledge, has a hierarchical structure that stresses individual achievement, and appears to maintain a fairly standardized and accessible set of rules governing classroom behavior and interaction"* (Treichler & Kramarae, 1983). When socio-technical practitioners acknowledge their participation in the systems of oppression and power and begin to interrogate themselves and their work as agents of privilege and power, there will be more critical engagement in AI ethics. Focusing on mathematical abstractions distracts from the socio-historical context needed to make a fairer world and removes accountability from researchers to place it on technology. A reasonable approach to tackle this issue could be to begin here. This is an introductory primer and critique of current AI ethics, and we present individual first steps and principles:

- Imagining solutions beyond reform alone that lead to organizing for systemic change

- Identifying and acknowledging systems of oppression
- Unpacking the way every person participates in these systems (unconsciously or consciously)
- Recognizing that human rights matter more than the rights and interests of organizations and corporations
- Further unpack the frame of systems of oppression by applying them to individual research as well as the larger research community
- At the beginning of every research endeavor, reflect on systemic and local impact to interrogate the work
- Examine (1) the impacting system of oppression, (2) current research interventions, and (3) the unseen negative impacts of performative AI ethics research without critical analysis

We implore researchers and tech practitioners to participate in abolitionist-framed AI ethics/fairness research, and to do so the very first step is to actively learn about systemic oppression and what that looks like for everyone. There is a moral obligation to engage with ethics beyond theoretical abstraction because the harmful consequences negatively impact real people. Technology and AI do not live in a vacuum void of systemic injustice or historical legacy; one must understand the real world to create tools that impact the real world, which begins with systems of oppression and how practitioners knowingly and unknowingly perpetuate them.

References

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., & Sadagopan, S. (2021). Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 329–341.

Aguilar, M. (2022). Why is it important to have women in tech: Ironhack blog..

Ahmed, S. (2004). Declarations of whiteness: The non-performativity of anti-racism. *borderlands*, *3*(2).

Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Machine bias..

Benjamin, R. (2020). Race after technology: Abolitionist tools for the new jim code..

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Binns, R., & Kirkham, R. (2021). How could equality and data protection law shape ai fairness for people with disabilities?. *ACM Transactions on Accessible Computing (TACCESS)*, *14*(3), 1–32.

Birch, K., & Bronson, K. (2022). Big tech..

Birhane, A., Ruane, E., Laurent, T., S. Brown, M., Flowers, J., Ventresque, A., & L. Dancy, C. (2022). The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, p. 948–958, New York, NY, USA. Association for Computing Machinery.

Bonilla-Silva, E. (2022). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. Rowman & Littlefield, an imprint of The Rowman & Littlefield Publishing Group, Inc.

Bouchrika, I. (2022). 75 u.s. college statistics: 2023 facts, data & trends..

Bufacchi, V. (2005). Two concepts of violence. *Political studies review*, *3*(2), 193–204.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.

Buyl, M., & De Bie, T. (2022). Inherent limitations of ai fairness. *arXiv preprint arXiv:2212.06495*.

Caldera, C. (2020). Fact check: Rates of white-on-white and black-on-black crime are similar..

Carey, A. N., & Wu, X. (2022). The fairness field guide: Perspectives from social and formal sciences. *arXiv preprint arXiv:2201.05216*.

Carr, L. G. (1997). *"Colorblind" Racism*. Sage.

Chin, C., & Robison, M. (2020). How ai bots and voice assistants reinforce gender bias. *Center for Technology Innovation. https://www. brookings. edu/research/how-ai-bots-and-voice-assistants-reinforce-gender-bias*.

Collins, P. H., & Bilge, S. (2020). *Intersectionality*. John Wiley & Sons.

Cooper, A. F., Moss, E., Laufer, B., & Nissenbaum, H. (2022). Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 864–876.

Costanza-Chock, S. (2018). Design justice, ai, and escape from the matrix of domination. *Journal of Design and Science*, *3*(5).

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. *University of Chicago Legal Forum*, *1989*(8), 139–167.

Crenshaw, K. (2017). Race liberalism and the deradicalization of racial reform. *Harvard Law Review*, 2298–2319.

Césaire, A. (1972). *Discourse on colonialism*. Monthly Review Press.

Davis, A. (2003). *Are Prisons Obsolete?* Seven Stories Press.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, *7*(7.4), 1.

Duvernay, A. (2016). 13th..

Dzieza, J. (2023). Ai is a lot of work..

Ehsan, U., Singh, R., Metcalf, J., & Riedl, M. (2022a). The algorithmic imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1305–1317.

Ehsan, U., Singh, R., Metcalf, J., & Riedl, M. (2022b). The algorithmic imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, p. 1305–1317, New York, NY, USA. Association for Computing Machinery.

Epatko, L. (2018). 5 years after the world's largest garment factory collapse, is safety in bangladesh any better?..

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Froehlich, T. J. (2017). A not-so-brief account of current information ethics: The ethics of ignorance, missing information, misinformation, disinformation and other forms of deception or incompetence.. *BiD*, *39*.

Gabriel, I. (2022). Toward a theory of justice for artificial intelligence. *Daedalus*, *151*(2), 218–231.

Gardner, J., & Warren, N. (2019). Learning from deep brain stimulation: the fallacy of techno-solutionism and the need for 'regimes of care'. *Medicine, Health Care and Philosophy*, *22*(3), 363–374.

Gebru, T. (2019). Oxford handbook on AI ethics book chapter on race and gender. *CoRR*, *abs/1908.06165*.

Gilmore, R. W. (2022). *Abolition Geography: Essays Towards Liberation*. Verso.

Google (2023). Violence..

Grant, C. (2022). Algorithms are making decisions about health care, which may only worsen medical racism: News & commentary..

Hagendorff, T. (2022). Blind spots in ai ethics. *AI and Ethics*, *2*(4), 851–867.

Hampton, L. M. (2021). Black feminist musings on algorithmic oppression. *arXiv preprint arXiv:2101.09869*.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 501–512, New York, NY, USA. Association for Computing Machinery.

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

Harro, B. (2000). The cycle of liberation. *Readings for diversity and social justice*, 52–58.

Hill Collins, P. (2000). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Hinton, E., Henderson, L., & Reed, C. (2018). For the record an unjust burden: The disparate treatment of black americans in the criminal justice system..

Hochschild, A. (2022). Leopold ii: King of belgium..

Hongladarom, S., & Bandasak, J. (2023). Non-western ai ethics guidelines: implications for intercultural ethics of technology. *AI & SOCIETY*, 1–14.

hooks, b. (1984). *Feminist Theory: From Margin to Center*. South End Press.

hooks, b. (2004). *The Will to Change: Men, Masculinity, and Love*. Washington Square Press.

Horvath, R. J. (1972). A definition of colonialism. *Current Anthropology*, *13*(1), 45–57.

Injustice, M. P. (2023). How many jurisdictions use each tool?. `https://pretrialrisk.com/national-landscape/how-many-jurisdictions-use-each-tool/`. Accessed: September 9, 2023.

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 310–323.

Johnson, K. (2020). Access now resigns from partnership on ai due to lack of change among tech companies..

Kara, S. (2018). Is your phone tainted by the misery of 35,000 children in congo's mines?..

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pp. 2564–2572. PMLR.

Keyes, O. (2018). The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, *2*(CSCW), 1–22.

Lechner, F. J., & Wallerstein, I. (2015). *The Modern World-System as a Capitalist World-Economy* (Fifth edition)., p. 56–62. Wiley Blackwell.

Lee, T. (2022). A deep dive into the labor exploitation behind everyday products..

Library, F. S. U. A. V. G.-C. (2023). Anti-racism resources: Colorblindness..

Loideain, N. N., & Adams, R. (2020). From alexa to siri and the gdpr: the gendering of virtual personal assistants and the role of data protection impact assessments. *Computer Law & Security Review*, *36*, 105366.

Lott, B. (2012). The social psychology of class and classism.. *American Psychologist*, *67*(8), 650.

Manne, K. (2017). *Down girl: The logic of misogyny*. Oxford University Press.

Marx, K. (1990). *Capital Volume 1*, pp. 11–57. Penguin Classics.

McKeever, E. (2022). Who turned out the lights? how critical race theory bans keep people in the dark. *Washington University Jurisprudence Review*, *15*(1).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Merriam-Webster (2023). Hegemonic definition..

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical ai. *Nature machine intelligence*, *1*(11), 501–507.

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*, 659–684.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of ai ethics. *Minds and Machines*, *31*(2), 239–256.

Morley, M. H. (1989). *The U.S. imperial state: theory and historical setting.* Cambridge University Press.

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism.* PublicAffairs.

Munn, L. (2023). The uselessness of ai ethics. *AI and Ethics*, *3*(3), 869–877.

NAACP (2021). The origins of modern day policing..

Nair, V. (2023). The environmental impact of llms..

Nazgol Ghandnoosh, P., & Nellis, A. (2022). Research - get the facts..

Noble, S. U. (2018). Algorithms of oppression. In *Algorithms of oppression.* New York University Press.

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), 447–453.

of Wisconsin Madison, U. (2023). Research guides: General philosophy: Ethics (moral philosophy) and value theory..

Oi, R. (2022). Women in tech: Pushing for inclusivity and diversity..

on Gender-Based Violence, A. P. I. (2018). Patriarchy & power..

O'neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown.

Panagiotopoulou, V.C. Stavropoulos, P., & Chryssolouris, G. (2022). A critical review on the environmental impact of manufacturing: a holistic perspective. *The International Journal of Advanced Manufacturing Technology*, 603–625.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect.* Basic books.

Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019). Meaningful explanations of black box ai decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9780–9784.

Perez, C. C. (2019). *Invisible women: Data bias in a world designed for men.* Abrams.

Perraudin, F. (2019). Low pay in the garment industry still a reality despite pledges – study..

Perrigo, B. (2023). Openai used kenyan workers on less than $2 per hour exclusive..

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. *Smart Health*, *20*, 100178.

Project, T. S. (2018). Report to the united nations on racial disparities in the u.s. criminal justice system..

Pulrang, A. (2022). 6 ways responses to covid-19 have been ableist, and why it matters..

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, p. 145–151, New York, NY, USA. Association for Computing Machinery.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–23.

Real, T. (2003). *How can I get through to you?: Closing the intimacy gap between men and women*. Simon & Schuster.

Report, A. I. (2023). Measuring trends in artificial intelligence. In *Artifical Intelligence Index Report 2023*. Stanford University Human Centered Artificial Intelligence.

Robinson, C. J., Sojoyner, D., & Willoughby-Herard, T. (1983). *Black Marxism, Revised and Updated Third Edition: The Making of the Black Radical Tradition* (3 edition)., pp. 75–77. University of North Carolina Press.

Rodriguez, D. (2020). Reformism isn't liberation, it's counterinsurgency..

Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, *2*(1). https://hdsr.mitpress.mit.edu/pub/7z10o269.

Schultz, M. D., & Seele, P. (2023). Towards ai ethics' institutionalization: knowledge bridges from business ethics to advance organizational ai ethics. *AI and Ethics*, *3*(1), 99–111.

Shanbehzadeh, M., Yazdani, A., Shafiee, M., & Kazemi-Arpanahi, H. (2022). Predictive modeling for covid-19 readmission risk using machine learning algorithms.. *BMC Medical Informatics & Decision Making*, *22*(1), 1–12.

Shew, A. (2020). Ableism, technoableism, and future ai. *IEEE Technology and Society Magazine*, *39*(1), 40–85.

Siapka, A. (2022). Towards a feminist metaethics of ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, p. 665–674, New York, NY, USA. Association for Computing Machinery.

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423*.

Smedley, A., & Smedley, B. D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race.. *The American psychologist*, *60 1*, 16–26.

Smith, L. (2023). Center for disability rights..

Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., Kuflik, T., & Otterbacher, J. (2019a). "end to end" towards a framework for reducing biases and promoting transparency of algorithmic systems. In *2019 14th International Workshop*

on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 1–6. IEEE.

Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., Kuflik, T., & Otterbacher, J. (2019b). "end to end" towards a framework for reducing biases and promoting transparency of algorithmic systems. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 1–6.

Tilmes, N. (2022). Disability, fairness, and algorithmic bias in ai recruitment. *Ethics and Information Technology*, *24*(2), 21.

Ton, J. (2020). Council post: Race in tech, part one: Inside the numbers..

Treichler, P. A., & Kramarae, C. (1983). Women's talk in the ivory tower. *Communication Quarterly*, *31*(2), 118–132.

van Maanen, G. (2022). Ai ethics, ethics washing, and the need to politicize data ethics. *Digital Society*, *1*(2), 9.

Vera (2023). Incarceration statistics..

Waelen, R. (2022). Why ai ethics is a critical theory. *Philosophy & Technology*, *35*(1), 9.

Wagner, W. S., & Peter (2022). Mass incarceration: The whole pie 2022..

Weinberg, L. (2022). Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ml fairness approaches. *Journal of Artificial Intelligence Research*, *74*, 75–109.

West, M., Kraut, R., & Ei Chew, H. (2019). I'd blush if i could: closing gender divides in digital skills through education..

Williams, R. M., Smarr, S., Prioleau, D., & Gilbert, J. E. (2022). Oh no, not another trolley! on the need for a co-liberative consciousness in cs pedagogy. *IEEE Transactions on Technology and Society*, *3*(1), 67–74.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, p. 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.