

Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness

Anaelia Ovalle
Department of Computer Science
University of California, Los Angeles

Arjun Subramonian
Department of Computer Science
University of California, Los Angeles

Vagrant Gautam
Spoken Language Systems
Saarland University

Gilbert Gee
Department of Community Health
University of California, Los Angeles

Kai-Wei Chang
Department of Computer Science
University of California, Los Angeles

ABSTRACT

Intersectionality is a critical framework that, through inquiry and praxis, allows us to examine how social inequalities persist through domains of structure and discipline. Given AI fairness' *raison d'être* of "fairness," we argue that adopting intersectionality as an analytical framework is pivotal to effectively operationalizing fairness. Through a critical review of how intersectionality is discussed in 30 papers from the AI fairness literature, we deductively and inductively: 1) map how intersectionality tenets operate within the AI fairness paradigm and 2) uncover gaps between the conceptualization and operationalization of intersectionality. We find that researchers overwhelmingly reduce intersectionality to optimizing for fairness metrics over demographic subgroups. They also fail to discuss their social context and when mentioning power, they mostly situate it only within the AI pipeline. We: 3) outline and assess the implications of these gaps for critical inquiry and praxis, and 4) provide actionable recommendations for AI fairness researchers to engage with intersectionality in their work by grounding it in AI epistemology.

CCS CONCEPTS

• **Social and professional topics** → **Computing / technology policy**; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

fairness, intersectionality, artificial intelligence, literature review

ACM Reference Format:

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2018. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 17 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00
<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Artificial intelligence (AI) fairness research is critical to the development of just AI. Work in this space consistently urges researchers and engineers alike to consider notions of fairness defined over model predictions. These notions vary across conceptualization (e.g., group, individual fairness [8]) and operationalization (e.g., pre/in/post-processing [3]) [53]; nevertheless, the literature generally agrees on the goal of minimizing negative outcomes across demographic groups, including groups associated with multiple, "intersectional" demographic attributes (e.g., Black women) [91]. However, Kong [64] observes that AI fairness papers often narrowly interpret intersectional subgroup fairness as intersectionality, the critical framework from which the term originates [29, 65]. This myopic conceptualization of intersectionality has non-trivial consequences for just AI design and epistemology (i.e., ways of knowing).

The term *intersectionality* describes a traveling framework of critical inquiry and praxis (i.e., practical action beyond mere academic theorizing) intended to examine interlocking mechanisms of structural oppression which produce inequality [29]. Critical inquiry into the formation of inequalities generates knowledge that can inform strategies for combating them, which is often referred to as praxis. Generating knowledge that illuminates the underlying mechanisms of oppressive systems is a shared objective among critical disciplines, such as feminist, antiracist, and decolonial studies, and is rooted in a history of resistance [28]. In science, knowledge production is inextricably tied to colonial legacies. Upon initial examination, science offers universal, empirically-grounded explanations for natural phenomena; however, science is rooted in colonialism through its imposing of a "a positivist paradigm¹ approach to research on the colonies and other oppressed groups" [21]. According to scientific colonialism, the researcher has "unlimited rights of access to source[s] of information belonging to [a] population," where data collection and knowledge formation reflects *the one reality* the researchers understands [20, 34]. Indigenous knowledge is erased as dominant knowledge systems are imposed, preventing Indigenous people from creating and sharing their own knowledge and perspectives. Consequently, disciplines rooted in colonial epistemology often assimilate prevailing knowledge systems that perpetuate the erasure of knowledge [21, 33, 42]. As such, critical disciplines do not decouple reclaiming knowledge from reclaiming power.

¹Knowledge as a result of "neutral" and quantifiable observation. This paradigm strictly relies on only measurement and reason[75].

The epistemologies of AI research are not divorced from scientific colonialism’s legacy. Intersectionality may be used to critically examine AI research methodologies, so that “the world-views of those who have suffered a long history of oppression and marginalization are given space to communicate from their frames of reference” [21]. Intersectionality promotes grappling with “how individuals and groups who are subordinated within varying systems of power might survive and resist their oppression,” thereby empowering communities to criticize the injustices they experience [28]. In the face of epistemic violence (e.g., the erasure of Indigenous knowledge), intersectionality erects a new form of epistemic resistance: knowledge production. Frameworks to articulate social inequalities have been integral to the survival of communities at the margins. Similarly, intersectionality, by enabling researchers to observe and articulate disparities, may break the epistemic molds “researchers are placed in so they may operate differently” [21].

In the context of AI fairness, intersectionality is less about getting technology right (e.g., establishing fairness constraints for a model); it is more about interrogating the social reality which drives AI oppression, so we can then make technology better. Crenshaw uses the term intersectionality as a metaphor to speak on how “different systems of oppression overlap,” but more importantly emphasizes that neglecting the convergence of these structures would cause rhetorical and identity politics to abandon issues and people who are actually affected by these intersecting “systems of subordination” [6]. Intersectionality thereby challenges the sociopolitical amnesia which frames subgroup fairness as solely a technical problem [91]. We do not reject subgroup fairness outright; rather, we share this example to challenge the AI fairness community to expand its engagement with intersectionality. To operationalize AI fairness with an intersectional lens, it is vital to first illuminate underexplored gaps between intersectionality and existing AI fairness literature. To this end, we ask: (1) how is intersectionality discussed in AI fairness literature?; (2) to what extent does this discussion change based on computer science (CS) methodology?; (3) where are the largest gaps in conceptualizing and operationalizing intersectionality for advancing social justice?; (4) what tensions exist in leveraging these gaps for just AI design?; and (5) what do these findings tell us about opportunities for more just AI? To answer these questions, we contribute the following:

- (1) Identify a growing body of AI fairness papers related to intersectionality (§4) and examine their conceptions of the critical framework in contrast to core intersectionality literature (§3).
- (2) Create guiding questions to critically assess the use of intersectionality as a lens to operationalize AI fairness (Table 2).
- (3) Use our findings to analyze where gaps remain in AI fairness papers’ use of intersectionality, and provide recommendations towards addressing these gaps (§5, §6). We further comment on the structural forces that may contribute to these observed norms in AI fairness research.

The majority of the papers we review approach intersectionality from the narrow perspective of subgroup fairness. Through a deductive lens in §5, we find that intersectionality engagement varies significantly depending on how it is situated within the AI pipeline, how sources of biases are described, and what CS research epistemologies are invoked. Inductively in §6, we find that even

when researchers center intersectionality literature, there is little engagement with the framework itself, evidenced by a lack of described social context, discussion of power and relations between structures, questionable citational practices, and a disjointed sense of social justice praxis.

Our paper does not concern itself with claiming that intersectionality must take a particular form within AI fairness. Rather, we center intersectionality as an “analytical sensibility” [22, 29], which when activated, can sharpen and transform the tools in the AI fairness researcher’s toolbox. This, we argue, is key to justice-centric AI development. We further seek to dispel the misconception that social science disciplines have no place in STEM [70, 77]. Educated in computer science, we equip the AI fairness researcher of similar training who is committed to justice with concrete ways of using their training in AI to exercise critical praxis. In this way, we hope to disrupt deep-rooted indifferences to social reality, “a powerful force that is perhaps more dangerous than malicious intent” [5].

Positionality Statement All but one author of this paper are formally trained primarily as computer scientists, with additional training in gender theory, critical social theories, criminology, linguistics, and related fields. One author is a social scientist who confronts issues of social inequities in both everyday life, and in their scholarship. This has led to a body of research that necessarily takes an intersectional and life course perspective. All authors have informal training in queer studies through activism and advocacy. As such, our backgrounds influence this work’s design, decisions, and development. All authors are located in the US or Europe, but have diasporic links to other social contexts; we do our best to position our work in a global context. We write this to empower individuals across both academia and industry research to critically engage with AI fairness paradigms. Therefore, our recommendations are articulated in a way that can be operationalized, though they are transferrable to other audiences. We position ourselves within a social justice ethos that is informed by decolonial theory, and that champions equity over equality as well as reparations to correct historical injustices.

2 RELATED WORKS

We are not the first to champion or critique intersectional praxis in AI fairness, let alone more broadly. Several works across disciplines including psychology and CS have advocated the use of intersectionality frameworks or discussed the misappropriation thereof [4, 14, 16, 24, 51, 55, 78, 80, *inter alia*]. Furthermore, AI ethics researchers have addressed the narrow perspective of intersectionality as intersectional subgroup fairness (e.g., Kong [64]); our review points to this too, although our scope is wider and considers numerous gaps in AI’s operationalization of intersectionality.

A few papers have reimagined intersectionality in AI [9, 23, 86], pushing for intersectional practices to be woven into the full AI pipeline, and arguing for a joint interrogation of culture, technology, and solutionist framings of fairness (e.g., critical technocultural discourse analysis [15]). Constanza-Chock [31] illuminates the lack of critical praxis in AI, drawing upon Collins’s matrix of domination to encourage researchers to reflect on how AI relates to “domination and resistance at each of these three levels (personal, community, and institutional)” [26]. Davis et al. [37], inspired by Crenshaw [36],

argue for AI to be reparative and aware of social and historical context. Klumbyte et al. [62] facilitate community-based critical analysis of the “tensions and possibilities” of integrating intersectional knowledge into machine learning systems. With a shared goal of intersectional AI, we complementarily gauge the epistemic alignment of AI papers related to intersectionality with Collins’ intersectionality tenets [29]. We go beyond the scope of papers like Birhane et al. [12], which is not explicitly about intersectionality and only assesses works with respect to their discussion of social context and power.

3 INTERSECTIONALITY OVERVIEW

Crenshaw coined the term “intersectionality” in her 1981 paper [35], and expanded on it in [36]. In the context of violence against Black women, these works study the interactions of race and gender, as well as racism and patriarchy as systems of subordination. Her work is grounded in “a bottom-up commitment” to address the needs of those who are “victimized by the interplay of numerous factors,” with the explicit goal of obtaining political and social justice. Thus, praxis has been an important facet of intersectionality from its inception; what constitutes praxis is broad and contextual, including “movements for economic justice, legal and policy advocacy, state-targeted movements for prison abolition” [22].

While various definitions of intersectionality have emerged, they all center a need to examine power relations across structures, disciplines, domains, and location [1, 25, 52]. We draw upon broad intersectionality scholarship in our paper to enrich our own observations. To ground our review methodology and analysis in the following sections, we base our evaluations on Collins and Bilge [29]. This work details six core tenets of intersectionality (drawing from an in-depth genealogy of intersectionality) that lend to an *analytical language* and *cognitive organization* around how forms of oppression are co-created, operated, amplified, and interact with social and structural disparities. These tenets are: social justice, social inequality, relationality, social power, social context, and complexity. We describe each tenet below, its connections to AI fairness, and how we interpret the tenet for advancing social justice in AI fairness. These tenets further inform the construction of 3-4 guiding questions per tenet to assess how well the works in our critical review engage with the tenets (Table 2).

Social Justice Intersectionality emerges as a synergy between inquiry and praxis, where praxis is action to advance social justice that is informed by inequities identified via critical inquiry (facilitated by the following tenets). Collins and Bilge [29] caution that inquiry alone does not further social justice; intersectionality “demands more than simply being critical and entails turning critical analyses into critical praxis” [29]. In AI fairness, social justice praxis spans numerous practical approaches to fairness, e.g., debiasing techniques, fairness metrics for multiplicative groups; however, its effectiveness depends on authors’ social context. Intersectionality widens these practical approaches; this does not remove researchers from the AI fairness domain, but rather deepens our ability to engage with the domain. Overall, intersectionality enables the creation of new forms of knowledge which are informed by a critical examination of how AI systems reproduce inequalities. Therefore, our social justice guiding questions assess how

works commit to advancing justice and center the perspectives of subordinated communities.

Social Inequality Intersectionality rejects the inevitability of inequality as “hardwired into the social world, into individual nature” [28]; rather, the framework emphasizes the study of how social inequalities are fundamentally formed and reinforced through saturated centers of power. Dismantling inequalities requires the locating of these centers. In AI fairness, inequality is often measured via quantities like demographic parity and disparate impact [53]. Hence, these metrics ground the practice of harm reduction; however, static measures pointing towards equality rather than equity do not resolve complex and wide-reaching inequality. Instead, intersectionality asks us to center the social and historical context of those at the margins to inform praxis. As such, our inequality guiding questions assess the depth with which researchers situate their work in social inequality.

Relationality Relationality enables us to observe structures via their relation to one another. Relationality comprises: addition (what happens when we *don’t* consider the intersections of social categories), articulation (how relations impact the growth or dissolution of such intersections), and co-formation (e.g., of social categories as phenomena) [29]. In the context of AI fairness, relationality involves examining the relations between decisions we make as researchers, the technical artifacts we produce, and whom they impact (e.g., how the Eurocentrism of auditing frameworks cause them to fail to capture inequalities in globally deployed AI). Hence, our relationality guiding questions assess works’ intention and inquiry across technological structures and social context.

Social Power Intersectionality uses relationality to tie “intersecting power relations” to how power “produce[s] social divisions of race, class, gender, class, etc.” [29]. Intersectionality is predicated on understanding that systems of power “co-produce each other in ways that reproduce unequal material outcomes and the distinctive social experiences [within] hierarchies” [28]. In AI fairness, power is concentrated in *human* choices: system design, data collection, deployment, operationalizations of fairness. These choices impact the allocation of resources for communities at the intersections of the “structural, disciplinary, cultural, and interpersonal” domains [2, 26, 31]; as such, power should be discussed at all stages of the AI pipeline. Thus, our power guiding questions assess the extent to which researchers reflexively comment or situate their work in the power relations in which they participate.

Social Context Intersectionality centers “context-specific [...] historical particularities and the increasing significance of a global context” [29]. When engaging with intersectionality in different (especially global) contexts, inquiry and praxis take different forms; consequently, one must practice epistemic, personal, and critical reflexivity to be cognizant of context, towards effectively and holistically advancing justice. In AI fairness, social context informs AI context through: researcher training and background, model training and deployment, language choices, etc. Hence, self-reflexively acknowledging that one operates in the Global North *informs* who is centered in fairness tasks. Conversely, fairness works that flatten social context (e.g., by optimizing for “Indigenous people” broadly) informs who drives knowledge production. As a result, our social context guiding questions assess the extent to which context is deliberately referenced and informs research processes.

Complexity Complexity is key to a “creative tension” between critical inquiry and praxis, which results in new forms of social action to combat inequality [29]. Complexity necessitates relational thinking and situational awareness. In AI fairness, complexity is often conceptualized as minimizing unfairness across a large number of social groups. However, complexity is more expansive; for example, it entails co-designing with groups who have been harmed by AI systems rather than using preconceptions of excluded groups to remedy exclusion. Our complexity guiding questions probe how works contend with model requirements, community needs, and centers of power that influence AI design.

4 CRITICAL REVIEW METHODOLOGY

4.1 Paper Inclusion Criteria

To gauge how AI fairness research conceptualizes and operationalizes intersectionality, we curate 30 papers by: 1) querying “intersectionality machine learning” on Google Scholar to obtain 75 relevant papers, and 2) filtering those to papers published in AI venues including symposiums, conferences, journals, and books. We choose to query “machine learning” as AI fairness research tends to center machine learning. Our process simulates how researchers might discover AI fairness literature related to intersectionality when grounding their own work. Papers are tagged as including intersectionality if they cite intersectionality scholarship that centers critical inquiry. We restrict our sample to 30 papers to ensure that we can annotate each paper (some papers by multiple authors) for engagement with intersectionality. We document all the papers we review in Tables 4 and 5, and provide statistics thereof in Table 1.

4.2 Review Methods

Our annotation scheme is based on the tenets and corresponding guiding questions discussed in §3. All questions reflect three axes of reflexivity: epistemological, personal, and critical [76]. For each paper, for each guiding question (e.g., “Do the authors mention power?”), we annotate whether or not the authors of the paper explicitly or implicitly answer the question. Then, for each tenet, we annotate that the paper has characteristics of the tenet if it explicitly or implicitly answers at least one of the guiding questions corresponding to the tenet. Importantly, our questions are not a checklist to determine whether researchers have “truly” engaged with intersectionality; rather, they reveal where efforts in AI fairness are concentrated and help us reimagine our practices towards advancing social justice in AI. We share all our guiding questions in Table 2. We further break down our methodology for creating questions in Appendix §B. Our annotations are available at <https://tinyurl.com/intersectionality-annotations>.

11 out of the 30 papers were evaluated by three annotators, and we present our tenet-level interannotator agreement for these papers in Table 3. The scores in Table 3 indicate moderate to high interannotator agreement. The remaining 19 papers were each evaluated by at least 1 annotator. We expand on our annotation methodology in Appendix §C.

Given the nature of intersectionality, engagement therewith cannot be captured solely through quantitative means; therefore, we also qualitatively mine intersectionality-related themes from our sample of papers. At the nexus of these deductive (i.e., using our

guiding questions) and inductive (i.e., qualitative coding) analyses, we supply a bird’s eye and granular view of engagement with intersectionality in AI fairness. As praxis, we translate our inductive findings to recommendations for deeper engagement with intersectionality. These recommendations are tailored for AI fairness researchers with any level of training in AI in academia, industry, or both. We urge readers to take their own identity, capacity, and power into account when considering our recommendations, as these will affect what they can do and potential consequences. In our analyses, we acknowledge that papers are products of varied epistemological contributions, relations between authors and reviewers, and power dynamics. Thus, our critical review is not so much a criticism of AI fairness researchers as it is a reflection of broader systems, such as the incentives and infrastructural forces that govern publishing in CS and enacting change in corporations, as well as the types of knowledge production that are valued or even simply considered legitimate in the field. Papers do not reflect everything that goes into a research project, and they are also merely static snapshots in time that researchers grow beyond.

4.3 Investigating Intersectionality Within the AI Fairness Research Paradigm

Reflexivity enables AI fairness researchers to engage in praxis; as Mohamed et al. [70] comment, “deciding what counts as valid knowledge, what is included within a dataset, and what is ignored and unquestioned, is a form of power [...] that cannot be left unacknowledged.” To interrogate knowledge and inspire reflexivity, we texture our deductive analysis of intersectionality in AI fairness via four methodology lenses: where intersectionality is situated in the AI development process, how papers describe sources of bias, types of CS papers, and disciplinary relationality (i.e., synergy). These methodologies speak to both the research process and structures which researchers navigate in their work. We document the methodology tags for all the papers we review in Table 4.

4.3.1 Operationalization of intersectionality. We observe how papers engage with and operationalize intersectionality in the AI pipeline. Papers are tagged as pre-, in-, and post-processing, full pipeline, or processes. “Full pipeline” situates intersectionality (for empirical work) across the pipeline, while “processes” situates intersectionality in broader AI design and epistemology. Works that deeply engage with intersectionality exercise its tenets at every stage of the pipeline. Researchers can contrast modes of operationalizing intersectionality and in that tension, reimagine how they engage with the framework.

4.3.2 Source of bias. Understanding sources of bias is pivotal to aligning AI fairness with intersectional praxis. Intersectionality posits that unequal outcomes reflect a systemic reproduction of existing power relations [28]. Hence, we observe the bias sources that authors use to motivate their paper. Systemic descriptions of bias concern structures and oppressive forces which subsequently permeate sociotechnical systems. In contrast, statistical descriptions limit sources of bias to the model or data. A paper may characterize bias as systemic, statistical, both, or entirely fail to describe its source. Knowing this enables researchers to situate their context and iterate on their problem formulation.

Table 1: Critical review statistics ($N = 30$)

Characteristic	N	%
Intersectionality literature referenced	22	0.73
No. papers for annotator agreement	11	0.37
Terminology		
Uses term “intersectionality”	26	0.87
Uses term “intersectional”	27	0.9
AI Pipeline Stage		
Pre-processing	5	0.17
In-processing	4	0.13
Post-processing	10	0.33
Full pipeline	5	0.17
Processes	10	0.33
Computer Science Research Paradigm		
Theoretical	10	0.33
Empirical	23	0.77
Engineering	11	0.37
Other	6	0.2
Synergy across disciplines	16	0.53

4.3.3 CS paper type. We study paper types considered valid in computer science, as determined by those in positions of power, which exposes tensions between intersectionality and visibility. This allows us to interrogate existing assumptions surrounding supposed barriers to knowledge due to disciplinary divides [77]. Therefore, we break down papers by theoretical, engineering, empirical, or a combination of types based on Stent [84]. Papers that do not fit any of these types are tagged as “other.” This information enables AI fairness researchers to interrogate possible interplays between intersectionality and their epistemology.

4.3.4 Synergy across disciplines. Inspired by decolonial theory, we illuminate disciplinary gaps conducive to intersectional praxis [81]. Smith [81] asserts that knowledge is always situated; dominant academic AI epistemologies describe systems as “universal” or “neutral,” when in fact these terms simply indicate that other ways of knowing have been subjugated. For AI fairness research, engaging in participatory AI is one way of “recovering [...] stories of the past” [88]. However, researchers can also embrace synergy across disciplines. This allows for room to both examine how AI epistemology’s alignment with other works interacts with intersectionality and makes room to reimagine new forms knowledge production towards advancing AI fairness. For our evaluation, papers are tagged for synergy if they incorporate literature beyond other AI papers and intersectionality scholarship (tagging process described in §A) which motivate their contributions. By incorporating knowledge forms beyond CS, we make room for dialogue across “more than one way of knowing” [81]. This is particularly important for sources of marginalized knowledge that may go unheard.

5 DEDUCTIVE ANALYSIS

Quantitative Summary. We report tenet distributions across all papers in Figure 1. Complexity (97% of all papers), inequality (83%), and justice (83%) appeared most often in works that engaged with at least 1 guiding question. In contrast, the tenets that appeared least often were power (53%), context (57%), and relationality (60%). Taking the number of questions answered as a proxy for depth of engagement with a tenet, we see drops in every tenet. The largest

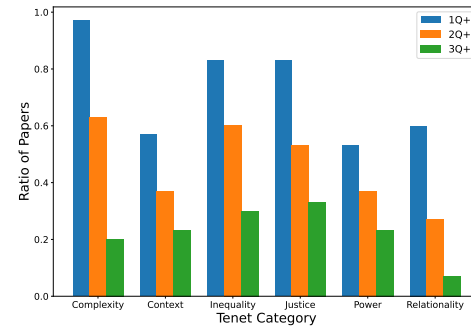


Figure 1: Distribution of intersectionality tenets split by depth of engagement, where 1Q+ indicates at least 1 characteristic was identified.

drop (20%) between answering 1+ questions versus 3+ questions is in complexity, despite high overall engagement. Relationality similarly drops from 60% to just 7%. These results are interrelated just as the tenets are; understanding power across structures requires understanding social context and the relations between social groups [30]. Therefore, it is suspect that a majority of papers purportedly center social justice and inequality with so few discussing power.

Has Intersectionality Citation. When examining how citation of intersectionality literature affects how papers engage with the tenets (Figure 2a), we find divergences in the consideration of power, inequality, and context. On the other hand, it does not seem to cause differential engagement with complexity and justice.

64% of papers that cite intersectionality engage with power, compared to 25% of papers that don’t cite it. As intersectionality is grounded in an analysis of power, engagement with the literature would explain this difference. However, the overall consideration of power is middling at best, echoing intersectionality theorists’ observation that the “recasting of intersectionality as a theory primarily fascinated with the infinite combinations and implications of overlapping identities from an analytic initially concerned with structures of power and exclusion is curious given the explicit references to structures that appear in much of the early work” [22].

We see a similarly large gap in engagement with inequality. 91% of papers that cite intersectionality also discuss social inequality as a phenomenon with social and historical roots, or something their work impacts, compared to only 62% of papers that don’t cite it. We see this difference as a reflection of intersectionality’s motivation as a framework to examine inequalities.

Papers only seem to show consistent engagement with the tenets of complexity and justice, regardless of whether they cited intersectionality literature (above 80% of papers in each of these splits). This reflects the CS paradigm of understanding intersectionality as rejecting single axes of identity, and the ethos of AI fairness – one that seeks justice and a better future. Overall, citing intersectionality literature correlates with deeper tenet engagement.

Operationalization of Intersectionality. Figure 2b shows differences in how intersectionality is used across the AI pipeline. Papers operationalizing intersectionality end-to-end had the largest

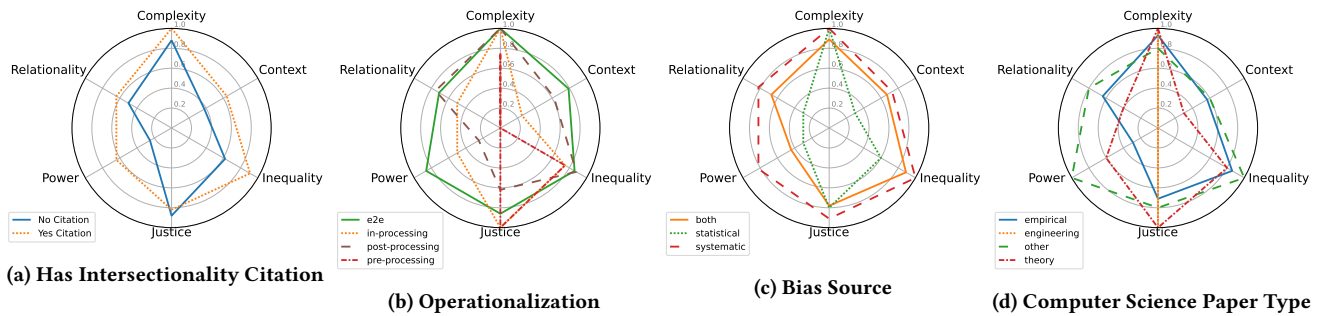


Figure 2: Relative distributions of papers across intersectionality tenets if a paper engages with at least 1 question per tenet.

coverage across intersectionality tenets, with each tenet appearing in 71-100% of these papers. Meanwhile, the lowest engagement across tenets came from papers focused on pre-processing, with *none* of them engaging with context, power and relationality.

The locus of operationalization seemed to make the biggest difference in how context and power were engaged with. Engagement with the social context tenet seemed to increase as papers went further down the AI pipeline; no pre-processing focused papers engaged with it compared to 25% of in-processing focused papers, 50% of post-processing focused papers, and 71% of end-to-end papers. This pattern mostly held for power as well, except that in-processing papers (50%) engaged with this tenet more than post-processing papers (25%). Overall, papers engage with more tenets when they operationalize intersectionality end-to-end and in processes.

Source of Bias. Differences in tenet engagement across bias source are shown in Figure 2c. Papers treating the source of bias as statistical had the lowest engagement across tenets, with only 30% of these papers engaging with context, relationality, and power. On the other hand, these papers have 100% coverage of the complexity tenet. This could be attributed to a common narrow reading of intersectionality as just multiplying identity categories rather than as a structural analysis or a political critique [50].

When considering bias to be systemic rather than statistical, tenet coverage increases noticeably; engagement with relationality jumps from 30% to 67% of papers in the category, context goes from 30% to 73%, and inequality goes from 60% to 100%. This aligns with existing literature in which discussing the social reality of a phenomenon allows one to more deeply assess the factors that contribute to it in the first place [5].

Papers that conceive of bias as *both* statistical and systemic have the best tenet coverage overall, with roughly 80-90% of papers discussing each of complexity, inequality, and justice. This dual conception of bias incorporates both the social and technical aspects of AI systems and how they may inform or magnify each other.

Computer Science Paper Type. The effect of CS paper types on tenet coverage is shown in Figure 2d. Papers across all types consistently engage with complexity and justice, with at least 70% of papers of each type covering these tenets. This consistency breaks down more dramatically across power, relationality, context, and inequality. *No* engineering papers engaged with these four tenets.

At the other end of the spectrum, papers classified as *other* engaged with the largest array of tenets. 100% of these papers engaged with power, as opposed to 60% of theory papers and a quarter of empirical papers. Theoretical papers seemed to engage relatively less with context (32%) and relationality (41%). Overall, despite disciplinary divides, papers in CS are able to engage with intersectionality tenets. Supplementing these findings, our inductive analysis in section §6 indicates that many works use a heuristic definition of intersectionality that is easily operationalized across theoretical, engineering, and empirical papers, resulting in a narrow use of the framework. Engaging with literature outside the empirical and engineering papers that are de rigeur in CS can expand tenet coverage.

Synergy Across Disciplines. Figure 3 shows each tenet category split by whether or not they had a synergistic component. As it pertains to bias source, synergistic papers incorporate a wider range of tenets at higher rates than non-synergistic papers (Figure 3a). Even among papers that treat bias as systemic and thus engage hugely from synergy, with 63-73% more papers discussing complexity and justice. Figure 3b shows that when papers that discuss intersectionality across the entire pipeline have a synergistic component, they have better tenet coverage. Synergy appears not to have a big effect on papers that focus on in-processing, pre-processing or post-processing, sometimes even appearing to decrease tenet coverage. With computer science paper types (Figure 3c), papers that incorporated intersectionality in an empirical and theoretical paradigm had better tenet coverage when they had a synergistic component. We note that no engineering papers in our data have a synergistic component – an interesting finding in its own right. As before, this suggests that the biggest benefit to tenet coverage can come from first operationalizing intersectionality throughout the pipeline and attending to processes and norms, which arguably *necessitates* interdisciplinary synergy. Overall, disciplinary synergy correlates with higher intersectionality tenet coverage.

6 INDUCTIVE ANALYSIS

Intersectionality as intersectional subgroup fairness. Among papers which cite intersectionality literature, many conflate intersectionality with intersectional subgroup fairness. For example, Fitzsimons et al. [44] claim:

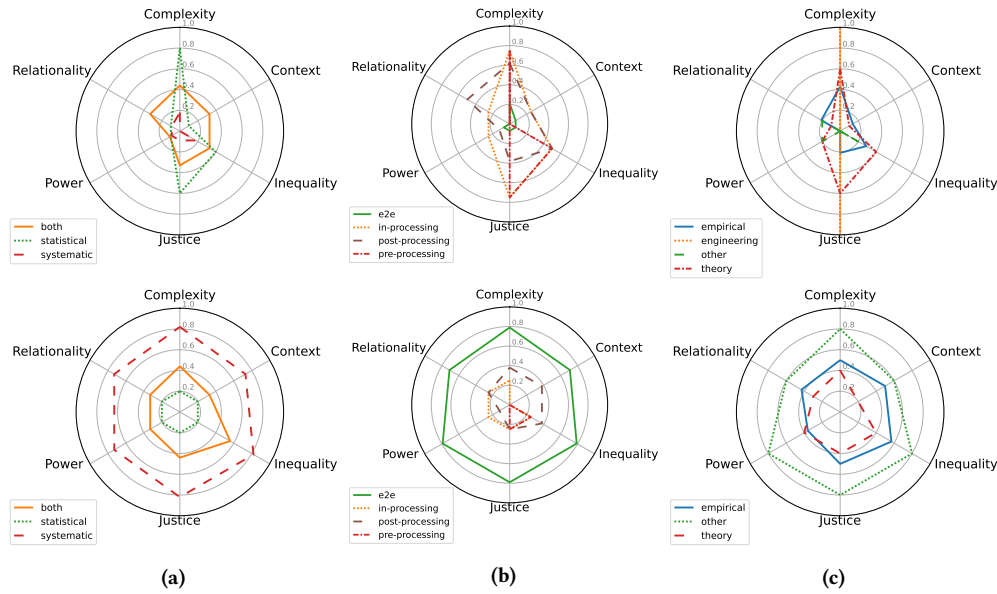


Figure 3: Papers with at least 1 tenet characteristic, split by presence of a synergistic literary component (top=no, bottom=yes).

“... a model that satisfies conditional parity with respect to race and gender independently may fail to satisfy conditional parity with respect to the conjunction of race and gender. In the social science literature concerns about, potentially discriminated against, sub-demographics are referred to as intersectionality.”

Similarly, Mougán et al. [72] state, “For intersectional fairness, we created the variable EthnicMarital, engineered by concatenating Ethnic and Marital status.” Indeed, we observe that many papers conceptualize intersectionality as identity-centric, and its ties to power and inequality are not explicitly named [71, 72, 93]. Our finding substantiates the concerns of intersectionality scholars that intersectionality is diluted to a “two-by-two analysis of gender by race” rather than “constituting a structural analysis or a political critique” [50], or contending with “overlapping systems of subordination” [30]. Notably, papers even discuss power and inequality in depth at first, but nevertheless operationalize intersectionality as subgroup fairness without engaging these points again [45].

Such additive frameworks are helpful insofar as they enable structural inquiry. However, per our annotations, despite overwhelming discourse on cross-sectional social categories, papers’ discussions of subgroups often lack social or historical context [18, 44, 71]. Few works comment on the structural factors that cause certain groups to be underrepresented in datasets, critically engage with the colonial origins of protected attributes [41], or connect groups to social structures and inequality (e.g., Black communities are targeted at a higher rate by law enforcement using facial recognition [17]).

Cho et al. [22] express confusion at this “recasting of intersectionality as a theory primarily fascinated with the infinite combinations and implications of overlapping identities from an analytic initially concerned with structures of power and exclusion,” despite sociotechnical systems encoding social relations and ideologies. The obfuscation of intersectionality as subgroup fairness reflects cultural denial, “the process that allows us to know about cruelty,

discrimination, and repression, but never openly acknowledge it” [40]. We do not claim that the intentions of AI fairness researchers are malicious; rather, groups “los[ing] meaning as a descriptive, non-analytical category” prevents researchers from engaging in critical inquiry [28]. This disarms praxis: AI fairness can no longer contend with advancing justice for those at the margins if their experiences with AI-driven social inequalities are not centered. Therefore, we echo Collins’ call for “intellectual vigilance” in analyzing and articulating intersecting power relations. Using an intersectional lens is crucial to refocusing on marginalized communities, and can inform social justice efforts across various fields by addressing the root causes of harm, regardless of one’s training.

Recommendation. Researchers exercise intellectual vigilance when using additive frameworks by creating statistical methodologies that preserve unique social and historical characteristics of intersecting groups. [90] exemplifies this. Leadership incentivizes this inquiry. Researchers and leadership prioritize widening their conceptualization of intersectionality beyond the “subgroup fairness” interpretation, which is limited in its social justice praxis.

Anti-discrimination legislation informs design. Several papers draw from regulation (e.g., anti-discriminatory legislation) to define their fairness objective. For example, Molina and Loiseau [71] state:

“In many—if not most—real-world applications, there are multiple protected attributes (typically 10-20) along which discrimination is prohibited [1, 2].”

Foulds and Pan [45] similarly motivate their fairness criteria from a legal perspective: “consider the 80% rule, established in the Code of Federal Regulation.” Additionally, Foulds et al. [46] seek to “[determine] whether disparities in system behavior meet legal thresholds for discrimination.” Furthermore, Ghosh et al. [48] remark, “there

does not exist a single universally agreed upon definition of fairness,” citing how different “anti-discrimination legislation exists in various jurisdictions around the world.”

Situating fairness in AI systems from a strictly regulatory lens (e.g., the 80% rule, protected groups) does not fully embrace social and historical context. Several critical scholars argue that discrimination is often legitimized through anti-discrimination law [35, 38, 47, 82]. According to Freeman [47], these laws see racial discrimination “not as conditions, but as actions inflicted on the victim by the perpetrator.” He adds that such laws reflect the idea that “only ‘intentional’ discrimination violates anti-discrimination principles,” creating “a class of the ‘innocent’ who need not feel any personal responsibility for the conditions associated with discrimination” [47]. Similarly, in AI fairness, researchers prioritize addressing intersectional subgroup fairness over the structures that give rise to unfairness to begin with. Interestingly, AI fairness researchers who adopt a regulatory lens abundantly cite Crenshaw [36], although this work illuminates how anti-discrimination laws render Black women invisible:

“In a monumental paper published in 1989, Kimberlé Crenshaw [11] introduced Intersectionality by referencing a court case where black women were unfairly discriminated as a result of an activity to mitigate the race and gender discrimination independently” [63].

AI fairness researchers must heed the warnings of critical legal studies: indifference to the social and historical context of groups and their intersections risks reproducing histories of discrimination. For example, an important step towards dismantling injustices is challenging social categories rooted in colonialism, as “this structure imports a descriptive and normative view of society that reinforces the status quo” [35]; hence, failing to investigate how the mechanisms responsible for the unjust social realities of oppressed groups are upheld by one’s technology is fundamentally incompatible with reparation and advancing justice [32, 37]. Ultimately, if the goal of researchers is to leverage intersectionality towards the creation of just AI systems, these systems must be infused with social and historical literacy throughout their lifecycle to prevent indifferent engagement with the people they affect.

Recommendation. Researchers, including those in leadership, critically engage with and remain vigilant of how operationalizations of anti-discrimination laws in their AI systems do not automatically mean that their systems are fair to marginalized communities. They may do this by engaging with critical legal studies texts [35, 38, 47, 82] and marginalized communities to learn how they are unfairly impacted even by systems that pass legal audits. [58] does a good job of examining the tensions between prioritizing different forms of fairness.

Angles of power examined: technodeterminism rules. Collins describes intersectionality as examining the mutual influences which “intersect and interlock” across “structural, disciplinary, cultural, and interpersonal” [26] domains of power. However, among papers that cite intersectionality, power is the least engaged tenet, with “power” mentioned in only 53% of papers. Moreover, merely mentioning “power” does not entail engaging with it in depth; for example, Foulds and Pan [45] write in their abstract that the “framework of intersectionality [...] analyzes how interlocking systems of

power and oppression affect individuals along overlapping dimensions,” but do not discuss power elsewhere in their paper. Similarly, Yang et al. [94] only mention power in their related works section.

Furthermore, across papers that do engage with power and power relations, engagement style varies. For example, we see power described as a distributable commodity; Suresh et al. [85] assert, “our work builds on intersectional feminist thought, and stems from the acknowledgement that power is not equally distributed in the world.” In contrast, Kirk et al. [60] note, without explicitly using the term “power,” that “models can exacerbate existing biases in data and perpetuate stereotypical associations to the harm of marginalized communities.”

One can argue that AI fairness researchers study mechanisms of inequality, namely the way inequalities emerge as “AI harms,” so that we may reduce them. As such, the allocational and representational harms of our systems are the result of power enacted by our systems unto those at the margins. However, many AI fairness researchers constrain their discussion of power to the AI system alone, removing themselves from the equation. The notion that a system itself exerts power is technodeterministic, i.e., it reifies the idea that systems, and not their creators, are responsible for reproducing inequalities. Only a few papers that we review escape technodeterminism, e.g., Kasy and Abebe [58] state, “The second alternative perspective focuses on the distribution of power and asks: who gets to pick the objective function of an algorithm? The choice of objective functions is intimately connected with the political economy question of who has ownership and control rights over data and algorithms.” Engaging with intersectionality forces researchers to shed their technodeterminism and contend with the value-laden choices made by the humans that contribute to the lifecycle of AI systems. This is central to praxis that may effectively advance justice in AI fairness.

Recommendation. Researchers flex intellectual vigilance by being explicit about how their methodologies may contribute to perpetuating social inequalities. They state their full-pipeline design choices at the beginning of projects and iterate as designs are updated. Leadership gives researchers opportunities to engage in critical reflexivity. These issues are further discussed in [5, 39, 73].

Questionable citational praxis of intersectionality. Several papers reference literature incorrectly to justify their operationalization of intersectionality. For example, Ghosh et al. [48] assume that Buolamwini and Gebru [17] concerns intersectionality (rather than intersecting subgroups), despite Buolamwini and Gebru [17] never citing any intersectionality literature themselves. We see this phenomenon again in Kang et al. [56], which cites only Buolamwini and Gebru [17] when describing intersectionality. In contrast, some papers, like Makhoul et al. [67], discuss intersectionality, but only cite a paper on affirmative action [57]. Other papers, like Foulds and Pan [45] and Mougán et al. [72], mention intersectionality, yet do not reference any relevant literature at all; this is reflected in our deductive analysis, with 19% of papers that use the term “intersectionality” not citing any intersectionality literature.

These findings exemplify a weak spot in the citational praxis of AI fairness researchers. Alexander-Floyd [1] calls for us to cite intersectionality research, showing that within social science literature, there has been an erasure of Black women and Black feminist

knowledge in papers that discuss intersectionality. She describes the centering of positivist and empiricist methods of knowledge production as a force that (re-)subjugates Black feminist knowledge and contributes to maintaining the status quo of whose knowledge counts as “scientific” [1]. Bilge [7] identifies similar power structures in feminist studies and the broader neoliberal academy that contribute to “neutralizing the critical potential of intersectionality for social justice-oriented change.”

We find this gentrification of intersectionality in our field too; AI research interprets intersectionality as a dimension of “solvability” and scale, and this “perpetuat[es] the status quo injustice” [64]. Furthermore, owing to a disciplinary gap, papers use vague language when describing intersectionality. For instance, Buolamwini and Gebru [17] assert, “the concept of Intersectionality covers diverse discussions including the issue of the oppression that people feel due to the discrimination [15].” Camara et al. [19] mention “the complex and interconnected nature of social biases.” Mitchell et al. [69] state, “an individual’s identity and experiences are shaped [...] by a complex combination of many factors.” Vague language prevents intersectionality from being appropriately situated in socio-technical systems. This may only impart a vague understanding of intersectionality to readers, neutralizing both the researchers’ and readers’ engagement with power structures and inequality.

Recommendation. Researchers explicitly share how their interpretation of intersectionality literature informs their methodology and assumptions. They read critical social justice literature outside of CS and cite it when incorporating it in AI design. Researchers, including those across leadership, expect and enforce intersectionality citational integrity when peer-reviewing.

Intersectional AI fairness lacks relationality. We find that AI fairness researchers have adopted intersectionality in a way that strips the relationship between structures from the complexity of intersectionality’s arguments. This “misrepresents [the] initial intent” of intersectionality [27, 92], i.e., to question “how larger social structures influence supposed group level differences” [16]. For instance, some works that engage with intersectionality literature propose statistical solutions for inequality. Foulds et al. [46] tackle data sparsity by exploiting the structure of data distributions of data-dense subgroups (e.g., white women, Black men) to inform the data distribution of data-sparse subgroups (e.g., Black women); they address that “it may be difficult [...] e.g. to estimate the overall behavior of a classifier on North American indigenous women who are immigrants to the USA, due to a lack of recorded data on such individuals,” and later remark, “if the mechanism’s behavior on women is informative of its behavior on Black women, we can accurately model all of the conditional probabilities with fewer parameters than empirical frequencies.” Molina and Loiseau [71] similarly seek to exploit structure: “Indeed, even if the protected attributes are independent, since the classifier makes predictions based on X which may encode redundant information from some A_k , there can be interaction between those protected attributes through the classifier.” We do not reject statistical approaches to reducing AI harms; however, formulations that do not situate their statistical methods in a social context by, for instance, stating statistical *and* social assumptions those methods are based on, entirely miss the point of intersectionality as a critical framework.

Being intellectually vigilant about the relationship between statistics and the social sciences is crucial for their intersection. However, different levels of contending with this intersection are observed. Vigilance is missing entirely when the assumptions and reasoning behind the translation from social science knowledge to statistics is not explicit (e.g., [18, 44, 72]), with Fitzsimons et al. [44] describing: “In the social science literature concerns about, potentially discriminated against, sub-demographics are referred to as intersectionality [12]. More formally, this work proposes a simple approach to ensure group fairness in expectation across an arbitrary set of subgroups.” Jin et al. [54] provides a more intentional socio-technical translation, “although all value combinations are assessed for intersectional fairness, some subgroups may be semantically meaningless and hence should not be returned as the output,” though what is “meaningful” is not described. Other works go into more depth with their assumption, (e.g., [90], [69]), with Mitchell et al. [69] stating with respect to subgroup formation that “collaboration with policy, privacy, and legal experts is necessary in order to ascertain which groups may be responsibly inferred, and how that information should be stored and accessed.”

We caution against citing intersectionality literature while ignoring the relationships between the structures that create social categories. This fortifies the fallacy that we have engaged in intersectional praxis if we statistically supplement missing knowledge without examining the embedded assumptions and implications of doing so. It is through this neutralization of critical vigilance and reflexivity that AI fairness researchers are unable to identify where social inequalities may emerge through their own praxis. Invoking an intersectional lens enables this and is, therefore, pivotal to understanding the interlocking systems that produce AI injustices and doing AI justice work.

Recommendation. Researchers engage in intellectual vigilance while translating social justice texts into research methodology. As a result, they preserve the social context of social groups when employing statistical methods, e.g., by transparently stating how they infuse statistical assumptions with context. Across points of power, researchers have “vigilance check-ins” to check translative assumptions during AI development milestones. [69] engages with transparency at the model level which complements these points.

AI social justice praxis varies. Some papers treat improved fairness as social justice praxis regardless of the task’s context. For example, Foulds et al. [46] use recidivism prediction as a fairness benchmark task. As recidivism prediction is a “byproduct of ongoing regimes of selective policing and punishment” [5], the task only serves to uphold the carceral state [51]. Here, intersectionality posits sites of violence are saturated intersections of power [29].

Furthermore, many works ignore social context, which ought to inform social justice praxis [59, 63, 79]. Some papers provide context (e.g., data collection is “biased toward non-minorities” [46]), but nevertheless prioritize generalization [46]. Some papers even give credence to inferring the social category of individuals; Fitzsimons et al. [44] state, “gender labels were inferred using the employees’ first names, parsed through the gender-guesser python library.” Furthermore, we identify works that highlight the oppressive nature of social categories though often defer contestations to future work. For example, Kirk et al. [60] advocate:

“Future research is recommended to make ground truth comparisons across a broader range of countries against the set of gender-intersections examined in this paper and to comment on a broader spectrum of gender identities.”

Moreover, few papers complement technical contributions with social action, and some even tout their “purely statistical approach” [71]; this neglects the complexity inherent to dismantling social injustices. Mathematical saviorism restricts the operationalization of critical praxis to the pre/in/post-processing stages. In doing so, it encourages AI researchers to locate sources of unfairness situated only within the technical domain, the broader sociotechnical milieu linked to the power relations and inequalities upheld by AI [39]. Consequently, people already at the margins are erased, even in these contexts that ostensibly address fairness, oppression, and complexity. Hence, AI fairness researchers must engage in praxis that is informed by the experiences of those at the margins.

Some papers justify design choices that do not center care for those at the margins through utilitarian perspectives, e.g., Molina and Loiseau [71] reason, “an algorithm which discriminates 1 person among a 1000 can be described as fair to an extent.” On the other hand, works like Suresh et al. [85] and Mitchell et al. [69] concretely advocate to dismantle injustice and shift power through participation in model development and transparency in deployment, respectively. The contrast in social justice praxis is notable; AI fairness researchers must consider how design choices situate AI systems within their sociotechnical context.

As Crenshaw [35] has said, “addressing the needs and problems of those who are most disadvantaged” means that “others who are singularly disadvantaged would also benefit.” Centering these people and the contexts tied to their oppression deepens social justice engagement and creates equity. This way of engaging with intersectionality thus equips AI fairness researchers, regardless of training, to better address inequalities and injustices in AI.

Recommendation. Researchers bridge social justice inquiry and praxis by investing in and valuing the knowledge from communities that their AI systems harm. Researchers and leadership make sure that the AI design process prioritizes harm reduction to promote justice for marginalized communities. [85] does a good job at centering AI development through community engagement.

AI fairness misses critical reflexivity. Several papers neglect to state their social context and its implications for research methodology. This is reflected in our annotations, with 43% of papers—even critical ones, e.g., Kong [64]—neglecting to state their social context (i.e., often the US) [18, 66, 93]. Furthermore, when describing social context, some works only include the US as an important context, without commenting on the aspects of complexity and power inherent to doing so. This privileges western contexts as the “default” context, resulting in western prototypicality (c.f., white prototypicality [49]). For instance, Kirk et al. [60] argue:

“using US data may provide an appropriate baseline comparison: 50% of Reddit traffic comes from the US, and a further 7% from Canada and the UK each [34]. Given that US sources form a majority in GPT-2’s training material [...], we consider the US dataset a satisfactory first benchmark.”

Moreover, when authors do name their social context, they often phrase it as a blanket limitation rather than a contextualization of their research choices; Yang et al. [94] share that “the social construction and definitions of sensitive attributes” are “outside the scope of the present work but which are important in any real application.” Stating their context as a limitation—instead of a point which textures their work from the onset—situates their context as an afterthought, rather than something that undergirds the entire research process. On the other hand, Suresh et al. [85] center reflexivity throughout their work stating: “Throughout this process, we take an explicitly feminist approach, both in our overarching process—which we strive to make iterative, reflexive, contextual, and participatory—as well as the technology we build”.

All in all, critical reflexivity is crucial to operationalizing intersectionality, both as inquiry and praxis. AI researchers are overwhelmingly located in the Global North [12], which makes many power relations and AI injustices invisible to us, especially when we lack the abilities to inquire upon it. Reflexivity requires that we observe the power relations we participate in or benefit from, dismantle these relations, and identify opportunities for social justice within AI fairness. Our advice aligns with conceptualizations of decolonization within the computational sciences; Birhane and Guest [11] comment that decolonizing “requires the beneficiaries of the current systems to acknowledge their privilege and actively challenge the system that benefits them.”

Works that decouple social context and relationality from intersectionality may reflect academic incentives (e.g., conference acceptances, funding [13], citations) and infrastructural forces (e.g., conference paper formats, objectivity-washing). These push AI researchers to make “fairness” palatable by treating it as a complexity-free scientific quantity that can be optimized [10, 87]. Our paper is bound by similar constraints; we empirically validate our critical analyses in order to publish and our citation of the papers we review gives them “academic currency” even as we critique them.

Recommendation. Researchers across points of power iteratively dialogue on unlearning “universal” frameworks of knowledge and remain vigilant of *whose* knowledge is centered when developing AI. Leadership incentivizes and provides resources for their team to engage in critical reflexivity tools throughout development. [61] provides a good example of iterative reflexivity.

7 CONCLUSION

What we cannot name, we cannot see. What we cannot see, we cannot address. By examining AI fairness papers related to intersectionality, we identify several patterns in how the literature discusses intersectionality and how it impacts our ability to produce equitable tools. While our field has much energy to get this technology right, we caution the community against mistaking surmounting a “fairness issue” pre/in/post the AI pipeline means we have fixed the social reality driving the problem. This work does not seek to discard existing AI fairness work; instead, we invite a widening of AI fairness practice by centering marginalized people and valorizing the critical knowledge production that make room for their voices. We provide recommendations grounded in producing critical knowledge on how AI systems reproduce social inequalities. Our recommendations are not mutually exclusive with respect to

AI fairness infrastructure. Rather, they empower researchers to flex the intellectual vigilance required to produce intersectional work, regardless of CS paradigm. Expanding both the conceptualization and operationalization of intersectionality will enable AI fairness researchers across points of power to engage in deeper social justice praxis for AI. To do this, we advocate for adopting intersectionality as an analytical sensibility rather than an axis of optimization.

*“I lack imagination you say
No. I lack language. The language to clarify
my resistance to the literate.”
- Cherríe Moraga (1983)*

REFERENCES

- [1] Nikol G. Alexander-Floyd. Disappearing acts: Reclaiming intersectionality in the social sciences in a post-black feminist era. *Feminist Formations*, 24(1):1–25, 2012. ISSN 2151-7371. doi: 10.1353/ff.2012.0003.
- [2] Solon Barocas and Andrew D. Selbst. Big data’s disparate impact. *California Law Review*, 104:671, 2016.
- [3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- [4] Greta R. Bauer, Mayuri Mahendran, Chantel Walwyn, and Mostafa Shokoohi. Latent variable and clustering methods in intersectionality research: systematic review of methods applications. *Social Psychiatry and Psychiatric Epidemiology*, 57:221–237, 2021.
- [5] Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social forces*, 2019.
- [6] Michele Tracy Berger and Kathleen Guidroz. *The intersectional approach: Transforming the academy through race, class, and gender*. Univ of North Carolina Press, 2010.
- [7] Sirma Bilge. Intersectionality undone: Saving intersectionality from feminist intersectionality studies. *Du Bois Review: Social Science Research on Race*, 10(2): 405–424, 2013. ISSN 1742-058X, 1742-0598. doi: 10.1017/S1742058X13000283.
- [8] Reuben Binns. On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2019.
- [9] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2): 100205, 2021. ISSN 2666-3899. doi: <https://doi.org/10.1016/j.patter.2021.100205>. URL <https://www.sciencedirect.com/science/article/pii/S2666389921000155>.
- [10] Abeba Birhane and Fred Cummins. Algorithmic injustices: Towards a relational ethics. *ArXiv*, abs/1912.07376, 2019.
- [11] Abeba Birhane and Olivia Guest. Towards decolonising computational sciences. *arXiv preprint arXiv:2009.14258*, 2020.
- [12] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. The forgotten margins of ai ethics. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 948–958, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533157. URL <https://doi.org/10.1145/3531146.3533157>.
- [13] Borhane Blili-Hamelin and Leif Hancox-Li. Making intelligence: Ethical values in iq and ml benchmarks. 2022.
- [14] Lisa Bowleg. Evolving intersectionality within public health: From analysis to action. *American journal of public health*, 111 1:88–90, 2021.
- [15] André Brock. Critical technocultural discourse analysis. *New Media & Society*, 20:1012–1030, 2018.
- [16] Nicole T. Buchanan, Desdamona Rios, and Kim A. Case. Intersectional cultural humility: Aligning critical inquiry with critical praxis in psychology. *Women & Therapy*, 43:235–243, 2020.
- [17] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*, 2018.
- [18] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie H. Morgenstern, and Duen Horng Chau. Fairvis: Visual analytics for discovering intersectional bias in machine learning. *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 46–56, 2019.
- [19] Antonio Camara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard S. Zemel. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in english, spanish, and arabic. In *LTEDI*, 2022.
- [20] Graham Cameron. Evidence in an indigenous world. In *Australasian Evaluation Society 2004 International Conference, Adelaide, South Australia*, 2004.
- [21] Bagele Chilisa. *Indigenous research methodologies*. Sage publications, 2019.
- [22] Sumi Cho, Kimberlé Williams Crenshaw, and Leslie McCall. Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of Women in Culture and Society*, 38(4):785–810, Jun 2013. ISSN 0097-9740, 1545-6943. doi: 10.1086/669608.
- [23] Sarah Ciston. Imagining intersectional ai. *xCoAx*, page 39, 2019.
- [24] Elizabeth R Cole. Intersectionality and research in psychology. *The American psychologist*, 64 3:170–80, 2009.
- [25] The Combahee River Collective. A black feminist statement. *Women’s Studies Quarterly*, 1978.
- [26] Patricia Hill Collins. Black feminist thought in the matrix of domination. 2000.
- [27] Patricia Hill Collins. Intersectionality’s definitional dilemmas. *Review of Sociology*, 41:1–20, 2015.
- [28] Patricia Hill Collins. *Intersectionality as critical social theory*. Duke University Press, 2019.
- [29] Patricia Hill Collins and Sirma Bilge. *Intersectionality*. John Wiley & Sons, 2020.
- [30] Patricia Hill Collins, Elaine Cristina Gonzaga da Silva, Emek Ergun, Inger Furseth, Kanisha D. Bond, and Jone Martínez-Palacios. Intersectionality as critical social theory. *Contemporary Political Theory*, 20:690–725, 2019.
- [31] Sasha Constanza-Chock. Introduction: #travelingwhiletrans, design justice, and escape from the matrix of domination. *Design Justice*, 2020.
- [32] A. Feder Cooper, Ellen Abrams, and NA Na. Emergent unfairness in algorithmic fairness-accuracy trade-off research. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.
- [33] Ruth Schwartz Cowan. Francis galton’s statistical ideas: the influence of eugenics. *Isis*, 63(4):509–528, 1972.
- [34] Fiona Cram. Kaupapa māori evaluation: Theories, practices, models, analyses. *Evaluation Hui Summit*, pages 11–16, 2004.
- [35] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge, 1989.
- [36] Kimberlé Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241, Jul 1991. ISSN 00389765. doi: 10.2307/1229039.
- [37] Jenny L. Davis, Apryl A. Williams, and Michael W. Yang. Algorithmic reparation. *Big Data & Society*, 8, 2021.
- [38] Richard Delgado and Jean Stefancic. *Critical race theory: An introduction*, volume 87. NYU press, 2023.
- [39] Catherine D’Ignazio. Data feminism: Teaching and learning for justice. *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, 2021.
- [40] Virginia E. Eubanks. Automating inequality: How high-tech tools profile, police, and punish the poor. 2018.
- [41] Frantz Fanon. Black skin, white masks. *My Black Stars*, 1952.
- [42] Frantz Fanon. The wretched of the earth. 1961. *Trans. Richard Philcox*. New York: Grove Press, 6, 2004.
- [43] Jessica Finocchiaro, Roland Maio, Faidra Georgia Monachou, Gourab K. Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. Bridging machine learning and mechanism design towards algorithmic fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2020.
- [44] Jack K. Fitzsimons, Michael A. Osborne, and Stephen J. Roberts. Intersectionality: Multiple group fairness in expectation constraints. *ArXiv*, abs/1811.09960, 2018.
- [45] James R. Foulds and Shimei Pan. An intersectional definition of fairness. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2018.
- [46] James R. Foulds, Rashidul Islam, Kamrun Keya, and Shimei Pan. Bayesian modeling of intersectional fairness: The variance of bias. In *SDM*, 2018.
- [47] Alan David Freeman. Legitimizing racial discrimination through antidiscrimination law: A critical review of supreme court doctrine. *Minn. L. Rev.*, 62:1049, 1977.
- [48] A. Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *AIDBEI*, 2021.
- [49] Lewis R. Gordon. *Is the Human a Teleological Suspension of Man? Phenomenological Exploration of Sylvia Wynter’s Fanonian and Biodecan Reflections*. Ian Randle, Kingston, Jamaica, 2006. ISBN 978-976-637-741-0.
- [50] Kathleen Guidroz and Michele Tracy Berger. A conversation with founding scholars of intersectionality.
- [51] Lelia Hampton. Black feminist musings on algorithmic oppression. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.
- [52] Ange-Marie Hancock. When multiplication doesn’t equal quick addition: Examining intersectionality as a research paradigm. *Perspectives on Politics*, 5(01), Mar 2007. ISSN 1537-5927, 1541-0986. doi: 10.1017/S1537592707070065. URL http://www.journals.cambridge.org/abstract_S1537592707070065.
- [53] Abigail Z. Jacobs and Hanna M. Wallach. Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2019.
- [54] Zhongjun (Mark) Jin, Mengjing Xu, Chenkai Sun, Abolfazl Asudeh, and H. V. Jagadish. Mithracoverage: A system for investigating population bias for intersectional fairness. *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020.
- [55] Seth C. Kalichman, Bruno Shkembi, and Lisa A. Eaton. Finding the right angle: A geometric approach to measuring intersectional hiv stigma. *AIDS and Behavior*, 26:27–38, 2021.
- [56] Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. Infofair: Information-theoretic intersectional fairness. *2022 IEEE International Conference on Big Data (Big Data)*, pages 1455–1464, 2021.
- [57] Sampath Kannan, Aaron Roth, and Juba Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248, 2019.
- [58] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.
- [59] Jae-Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. Intersectional bias in hate speech and abusive language datasets. *ArXiv*, abs/2005.05921, 2020.
- [60] Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frédéric A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular

- generative language models. In *Neural Information Processing Systems*, 2021.
- [61] Goda Klumbyte, Claude Draude, and Alex S. Taylor. Critical tools for machine learning: Working with intersectional critical concepts in machine learning systems design. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.
- [62] Goda Klumbyte, Claude Draude, and Alex S. Taylor. Critical tools for machine learning: Working with intersectional critical concepts in machine learning systems design. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1528–1541, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533207. URL <https://doi.org/10.1145/3531146.3533207>.
- [63] Kenji Kobayashi and Yuri Nakao. One-vs.-one mitigation of intersectional bias: A general method to extend fairness-aware binary classification. *ArXiv*, abs/2010.13494, 2020.
- [64] Youjin Kong. Are “intersectionally fair” ai algorithms really fair to women of color? a philosophical analysis. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.
- [65] Youjin Kong. Are “intersectionally fair” ai algorithms really fair to women of color? a philosophical analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 485–494, 2022.
- [66] John P. Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and A. Abbasi. Benchmarking intersectional biases in nlp. In *North American Chapter of the Association for Computational Linguistics*, 2022.
- [67] Karima Makhoul, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter*, 23: 14–23, 2021.
- [68] Vishwami Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3:659–666, 2021.
- [69] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.
- [70] Shakir Mohamed, Marie-Therese Png, and William Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659–684, 2020.
- [71] Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. *ArXiv*, abs/2206.05828, 2022.
- [72] Carlos Mougán, José Manuel Álvarez, Gourab K. Patro, Salvatore Ruggieri, and Steffen Staab. Fairness implications of encoding protected categorical attributes. *ArXiv*, abs/2201.11358, 2022.
- [73] Safiya Umoja Noble. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, New York, 2018. ISBN 978-1-4798-4994-9.
- [74] Lorelli S. Nowell, Jill M. Norris, Deborah E. White, and Nancy J. Moules. Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1):1609406917733847, 2017. doi: 10.1177/1609406917733847. URL <https://doi.org/10.1177/1609406917733847>.
- [75] University of Nottingham. Understanding Pragmatic Research — nottingham.ac.uk. <https://www.nottingham.ac.uk/helmopen/rlos/research-evidence-based-practice/designing-research/types-of-study/understanding-pragmatic-research/section02.html#:~:text=Positivism%20is%20a%20paradigm%20that,cannot%20be%20known%20for%20certain>. [Accessed 15-Mar-2023].
- [76] Erlinda C Palaganas, Marian C Sanchez, Visitacion P Molintas, and Ruel D Carcavito. Reflexivity in qualitative research: A journey of learning. *Qualitative Report*, 22(2), 2017.
- [77] Inioluwa Deborah Raji, Morgan Klaus Scheuerman, and Razvan Amironesei. You can't sit with us: Exclusionary pedagogy in ai ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 515–525, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445914. URL <https://doi.org/10.1145/3442188.3445914>.
- [78] Yolanda A. Rankin, Jakita Owensby Thomas, and Nicole M. Joseph. Intersectionality in hci. *Interactions*, 27:68–71, 2020.
- [79] Kenneth S. Rogerson and Aidan Fitzsimmons. Intersectional identities and machine learning: Illuminating language biases in twitter algorithms. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2022.
- [80] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. Intersectional hci: Engaging identity through gender, race, and class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 5412–5427, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346559. doi: 10.1145/3025453.3025766. URL <https://doi.org/10.1145/3025453.3025766>.
- [81] Linda Tuhiwai Smith. *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing, 2021.
- [82] Dean Spade. *Normal life: Administrative violence, critical trans politics, and the limits of law*. Duke University Press, 2015.
- [83] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2020.
- [84] Amanda Stent. *howtoreadadpaper.pdf*. <https://people.cs.pitt.edu/~litman/courses/cs2710/papers/howtoreadadpaper.pdf>, 2023. (Accessed on 03/06/2023).
- [85] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Araujo Cruxên, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. Towards intersectional feminist and participatory ml: A case study in supporting femicide counterdata collection. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.
- [86] Miriam E Sweeney and André Brock. Critical informatics: New methods and practices. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–8, 2014.
- [87] Zeerak Talat, Joachim Bingel, and Isabelle Augenstein. Disembodied machine learning: On the illusion of objectivity in nlp. *ArXiv*, abs/2101.11974, 2021.
- [88] Vivetha Thambinathan and Elizabeth Anne Kinsella. Decolonizing methodologies in qualitative research: Creating spaces for transformative praxis. *International Journal of Qualitative Methods*, 20:16094069211014766, 2021.
- [89] Sandhya Tripathi, Bradley A. Fritz, Michael S. Avidan, Yixin Chen, and Christopher R. King. Algorithmic bias in machine learning based delirium prediction. *ArXiv*, abs/2211.04442, 2022.
- [90] Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.
- [91] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 336–349, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533101. URL <https://doi.org/10.1145/3531146.3533101>.
- [92] Zeerak Waseem, Joachim Bingel, and Isabelle Augenstein. Disembodied machine learning: On the illusion of objectivity in nlp. *ArXiv*, abs/2101.11974, 2021.
- [93] Ke Yang, Biao Huang, Julia Stoyanovich, and Sebastian Schelter. Fairness-aware instrumentation of preprocessing pipelines for machine learning. 2020.
- [94] Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal intersectionality for fair ranking. *ArXiv*, abs/2006.08688, 2020.

APPENDIX

A TAGGING FOR INTERSECTIONALITY LITERATURE

Initially, we only tagged works as incorporating intersectionality literature when they included Collins and Bilge [29], Crenshaw [35], Hancock [52], or Cho et al. [22]. However, during our weekly discussions, we noticed that papers cited a wider array of intersectionality works; either other works by these same authors, or other scholars who center intersectionality within critical disciplines. Because we want to gauge how AI fairness conceptualizes intersectionality, casting a wider net on tags is valuable in that we can include works that, while unaware of our initial list of texts, state intersectionality as a motivation in their work and cite other works about intersectionality like [26], [30], [27], or [36]. As a result, we tag the presence of intersectionality literature if any paper includes works that: 1) discuss intersectionality outside of CS and 2) frames intersectionality as a critical social inquiry and praxis framework.

B GUIDING QUESTIONS AND CONSIDERATIONS

We chose creating 3-4 guiding questions to balance in-depth coverage of each tenet with annotation feasibility. We share all our guiding questions in Table 2. While some guiding questions are straightforward (e.g., “Do the authors consider cross-sectional social categories?”), others are more up to our interpretation and experiences (e.g., “Are there any discussions on how spaces operate at different domains of power?”). Our interpretation of the intersectionality tenets for advancing justice in AI fairness is influenced by our social context and location, including our formal AI training, social identities, and experienced social inequalities (§1). For instance, we (the investigators) are all trans and people of color, and hence were likely more attune to the discussion in Kong [64] of power differentials in “regular” experiences, e.g., going through airport security.

C ANNOTATION METHODOLOGY

We follow Lincoln and Guba’s 1981 model of trustworthiness in our analysis [74], taking steps to maximize its credibility, dependability, confirmability, and transferability.

- **Credibility:** We are highly familiar with Collins and Bilge’s tenets. We also engaged in in-depth intersectionality intensives hosted by Black feminist scholars. Furthermore, we all have done justice work in some capacity. The majority of authors on this paper are trans people of color operating in AI. One author is a social scientist who confronts social inequities in their scholarship through intersectional perspectives. We spent over 6 months developing the guiding questions.
- **Dependability:** 11 out of the 30 papers were evaluated by three annotators, and we present our tenet-level interannotator agreement for these papers in Table 3. The scores in Table 3 indicate moderate to high interannotator agreement. The remaining 19 papers were each evaluated by at least 1 annotator.

- **Confirmability:** During weekly investigator meetings, we discussed our guiding questions and identified major sources of disagreement in our annotations.
- **Transferability:** Our guiding questions can be operationalized across paper types and domains outside of academia.

D MEASURING INTERANNOTATOR AGREEMENT

We use Randolph’s κ to estimate inter-annotator agreement, which is free-marginal rather than fixed-marginal; we choose this because κ is computed over six distinct items (i.e., tenets).

Table 2: Collins and Bilge's tenets of intersectionality and our corresponding guiding questions

Tenet	Guiding Questions
Social inequality	<ol style="list-style-type: none"> 1) Do the authors ground their work in how specific social or historical contexts factor into social inequality? 2) Do the authors acknowledge the implications of their work with respect to social inequality? 3) Is there a discussion of how intersecting power relations produce social inequality?
Social power	<ol style="list-style-type: none"> 1) Do the authors mention power? 2) Do the authors discuss any movement of power to the powerless? 3) Do the authors mention the mutual construction of power? 4) Is their own power in the work named or do the authors reflexively comment on the oppressive power relations within which their work participates?
Social Context	<ol style="list-style-type: none"> 1) Do the authors name their social context or social location with respect to their work? 2) Do the authors discuss how their social context influences their ideas and work's design, decisions, and development? 3) Do they acknowledge the limitations of their contexts?
Relationality	<ol style="list-style-type: none"> 1) Do the authors discuss the relationships between either social groups or structures? 2) Do the authors engage with how different social groups, typically treated as separate, face shared oppression? 3) Do the authors comment on how their identities shape their inquiry in relation to the people affected by their work?
Complexity	<ol style="list-style-type: none"> 1) Do the authors consider cross-sectional social categories? 2) Do they involve those without power in the generation and social construction of new knowledge? 3) Do the authors comment on the interplay between technical interventions and social action, or critical inquiry and practice? 4) Are there any discussions on how spaces operate at different domains of power ?
Social justice	<ol style="list-style-type: none"> 1) Do the authors state their commitment or motivation as social justice? 2) Do the authors discuss ways in which fair predictions or rules are not equally applied to everyone and can still produce unfair and unequal outcomes? 3) Do authors aim to dismantle a form of injustice, rather than solely documenting it in the form of a paper?

Table 3: Tenet-level interannotator scores by Randolph's κ and % agreement

Paper	κ	% agree
Wang et al. [90]	1.0000	100.00
Foulds et al. [46]	1.0000	100.00
Kong [64]	0.7778	83.33
Foulds and Pan [45]	1.0000	100.00
Rogerson and Fitzsimmons [79]	0.5556	66.67
Kobayashi and Nakao [63]	0.5556	66.67
Kirk et al. [60]	0.5556	66.67
Buolamwini and Gebru [17]	1.0000	100.00
Ghosh et al. [48]	0.5556	66.67
Kasy and Abebe [58]	0.7778	83.33
Molina and Loiseau [71]	0.7778	83.33
Average:	0.7778	83.33

Table 4: Papers with AI fairness research methodology tags

ID	Paper	Bias Source	Intersectionality Operationalization	CS Paper Type	Synergy
1	Wang et al. [90]	statistical	full pipeline	empirical	yes
2	Foulds et al. [46]	statistical	in-processing	theoretical, engineering, empirical	no
3	Kong [64]	systemic	processes	other	yes
4	Lalor et al. [66]	both	post-processing	empirical	no
5	Foulds and Pan [45]	both	in-processing	theoretical, engineering, empirical	yes
6	Rogerson and Fitzsimmons [79]	systemic	post-processing	empirical	yes
7	Suresh et al. [85]	systemic	processes, full pipeline	empirical	yes
8	Klumbyte et al. [61]	systemic	processes	empirical, other	yes
9	Kobayashi and Nakao [63]	statistical	full pipeline	engineering, empirical	no
10	Kirk et al. [60]	both	post-processing	empirical	no
11	Kim et al. [59]	both	post-processing	empirical	no
12	Yang et al. [94]	both	full pipeline	theoretical, engineering, empirical	no
13	Buolamwini and Gebru [17]	statistical	pre-processing, processes	engineering, empirical	yes
14	Fitzsimons et al. [44]	statistical	in-processing	theoretical, engineering, empirical	no
15	Ghosh et al. [48]	systemic	post-processing, processes	theoretical, empirical	yes
16	Davis et al. [37]	systemic	processes	theoretical	yes
17	Steed and Caliskan [83]	both	post-processing	empirical	yes
18	Mitchell et al. [69]	systemic	processes	other	yes
19	Kasy and Abebe [58]	systemic	post-processing, processes	theoretical, empirical	yes
20	Cabrera et al. [18]	statistical	post-processing	engineering, empirical	no
21	Kang et al. [56]	statistical	in-processing	theoretical, engineering, empirical	no
22	Jin et al. [54]	statistical	pre-processing	engineering	no
23	Mhasawade et al. [68]	both	processes	other	yes
24	Camara et al. [19]	both	post-processing	empirical	yes
25	Yang et al. [93]	statistical	full pipeline	engineering, empirical	no
26	Molina and Loiseau [71]	statistical	pre-processing	theoretical, engineering, empirical	no
27	Tripathi et al. [89]	both	pre-processing	empirical	yes
28	Mougán et al. [72]	systemic	pre-processing	theoretical, empirical	no
29	Finocchiaro et al. [43]	systemic	processes	other	yes
30	Makhlouf et al. [67]	systemic	post-processing	other	no

Table 5: Papers with intersectionality-related reference tags

ID	Paper	Cites Intersectionality Literature	Says “Intersectional”	Says “Intersectionality”
1	Wang et al. [90]	Yes	Yes	Yes
2	Foulds et al. [46]	Yes	Yes	Yes
3	Kong [64]	Yes	Yes	Yes
4	Lalor et al. [66]	No	Yes	Yes
5	Foulds and Pan [45]	Yes	Yes	Yes
6	Rogerson and Fitzsimmons [79]	Yes	Yes	Yes
7	Suresh et al. [85]	Yes	Yes	Yes
8	Klumbyte et al. [61]	Yes	Yes	Yes
9	Kobayashi and Nakao [63]	Yes	Yes	Yes
10	Kirk et al. [60]	Yes	Yes	Yes
11	Kim et al. [59]	No	Yes	No
12	Yang et al. [94]	Yes	Yes	Yes
13	Buolamwini and Gebru [17]	No	Yes	Yes
14	Fitzsimons et al. [44]	Yes	Yes	Yes
15	Ghosh et al. [48]	Yes	Yes	Yes
16	Davis et al. [37]	Yes	Yes	Yes
17	Steed and Caliskan [83]	Yes	Yes	Yes
18	Mitchell et al. [69]	Yes	Yes	Yes
19	Kasy and Abebe [58]	Yes	Yes	No
20	Cabrera et al. [18]	No	Yes	Yes
21	Kang et al. [56]	No	Yes	No
22	Jin et al. [54]	No	Yes	No
23	Mhasawade et al. [68]	Yes	No	Yes
24	Camara et al. [19]	Yes	Yes	Yes
25	Yang et al. [93]	Yes	Yes	Yes
26	Molina and Loiseau [71]	Yes	Yes	Yes
27	Tripathi et al. [89]	No	No	Yes
28	Mougán et al. [72]	Yes	Yes	Yes
29	Finocchiaro et al. [43]	No	Yes	Yes
30	Makhlouf et al. [67]	Yes	No	Yes