The Forgotten Margins of AI Ethics

ABEBA BIRHANE, Mozilla Foundation & School of Computer Science, University College Dublin, Ireland
ELAYNE RUANE, SFI Lero & School of Computer Science, University College Dublin, Ireland
THOMAS LAURENT, SFI Lero & School of Computer Science, University College Dublin, Ireland
MATTHEW S. BROWN, Department of Computer Science, Bucknell University, USA
JOHNATHAN FLOWERS,,
ANTHONY VENTRESQUE, SFI Lero & School of Computer Science, University College Dublin, Ireland
CHRISTOPHER L. DANCY, Dept of Industrial and Manufacturing Engineering & Dept of Computer Science and Engineering, The Pennsylvania State University, USA

How has recent AI Ethics literature addressed topics such as fairness and justice in the context of continued social and structural power asymmetries? We trace both the historical roots and current landmark work that have been shaping the field and categorize these works under three broad umbrellas: (i) those grounded in Western canonical philosophy, (ii) mathematical and statistical methods, and (iii) those emerging from critical data/algorithm/information studies. We also survey the field and explore emerging trends by examining the rapidly growing body of literature that falls under the broad umbrella of AI Ethics. To that end, we read and annotated peer-reviewed papers published over the past four years in two premier conferences; FAccT and AIES. We organize the literature based on an annotation scheme we developed according to three main dimensions: whether the paper deals with concrete applications, use-cases, and/or people's lived experience; to what extent it addresses harmed, threatened, or otherwise marginalized groups; and if so, whether it explicitly names such groups. We note that although the goals of the majority of FAccT and AIES papers were often commendable, their consideration of the negative impacts of AI on traditionally marginalized groups remained shallow. Taken together, our conceptual analysis and the data from annotated papers indicate that the field would benefit from an increased focus on ethical analysis grounded in concrete use-cases, people's experiences, and applications that as well as approaches that are sensitive to structural and historical power asymmetries.

Additional Key Words and Phrases: ???

ACM Reference Format:
Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The Forgotten Margins of AI Ethics. 1, 1 (April 2022), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Abeba Birhane, abeba@mozillafoundation.org, Mozilla Foundation & School of Computer Science, University College Dublin, Belfied, Dublin, Ireland; Elayne Ruane, elayne.ruane@ucdconnect.ie, SFI Lero & School of Computer Science, University College Dublin, Belfied, Dublin, Ireland; Thomas Laurent, thomas.laurent@ucd.ie, SFI Lero & School of Computer Science, University College Dublin, Belfied, Dublin, Ireland; Matthew S. Brown, msb027@bucknell.edu, Department of Computer Science, Bucknell University, , , , USA, ; Johnathan Flowers, j.charlesflowers@gmail.com, , , , , , ; Anthony Ventresque, anthony.ventresque@ucd.ie, SFI Lero & School of Computer Science, University College Dublin, Belfied, Dublin, Ireland; Christopher L. Dancy, cdancy@psu.edu, Dept of Industrial and Manufacturing Engineering & Dept of Computer Science and Engineering, The Pennsylvania State University, , University Park, PA, USA, .

1 INTRODUCTION

The urgency and profound importance of fairness, justice, and ethics in AI has been made evident by grave failures, limitations, harms, threats, and inaccuracies emanating from algorithmic systems deployed in various domains from criminal justice systems, and education, to medicine. These have been brought forth by landmark studies [5, 11, 57] and foundational books [15, 20, 23, 27, 55, 58, 59, 84]. Furthermore, the critical importance of the topic is marked by newly founded conferences exclusively dedicated to AI Ethics (e.g. AIES and FAccT), the fast-growing adoption of ethics into syllabi in computational departments [24], newly introduced requirement for the inclusion of broader impacts statements for ML papers submitted in premier AI conferences such as NeurIPS[1], increased attention to policy and regulation of AI [36], and increasing interest in ethics boards and research teams dedicated to ethics in major tech corporations.

This urgency and heightened attention on fairness, justice, and ethics is warranted. AI research, tools, and models — especially those integrated into the social sphere — are not purely technical systems that remain in the lab but socio-technical systems that are often deployed into the social world with potential and real impact on individual people, communities, and society at large. From discriminatory policing [10, 46, 66] and discriminatory medical care allocations [57, 60, 78], to discriminatory targeted ads [3, 44, 69], algorithmic systems present real harm to the most marginalized groups of society and furthermore pose a threat to human welfare, freedom, and privacy.

The broadly construed field of AI Ethics has been crucial in bringing the societal implications of algorithmic systems to the fore. It is important to note that critical analysis of computing has existed as long as the computing and AI fields themselves [1, 21, 80, 82]. Nonetheless, over the past few years a robust body of work has established the field of AI Ethics as a foundational and essential endeavour within the broad AI and society research. Work on fairness, accountability, transparency, explainability, bias, equity, and justice are generally conceived under this broad umbrella concept of AI Ethics.

With algorithmic bias, harm, discrimination, and injustice in the spotlight, the broadly construed field of AI Ethics is growing rapidly, producing a large and diverse body of work over the past number of years. This growing field, far from representing a unified entity, is marked by plural, overlapping, competing, and at times contradictory values, objectives, methods, perspectives, and goals. Yet, despite the plurality, the emerging field occupies a critical role in setting standards and norms for AI ethics research and practice. From informing policy and establishing ethical standards, to (directly or indirectly) impacting funding and influencing research agendas, as well as shaping the general discourse of AI Ethics, this fast expanding field impacts the current and future direction of fairness, justice, and ethics research and practice in AI.

Historically, within the Western context, "classic" ethics, as a branch of philosophy has mainly been speculative and theoretical where queries of AI ethics emanating from them can veer towards hypothetical questions such as "threats from super-intelligence". While theoretical rigour is crucial, it also poses a risk where questions of AI ethics could get lost in abstract speculation at the expense of currently existing harms to actual people. As such, it is important to review the historical roots of the field, examine the limitations of "classic" ethics, understand the current state and emerging patterns of the field, explore where it might be heading, and provide recommendation towards AI ethics research and practice that is grounded in the concrete. To that end, in this paper we:

- Trace historical roots of current AI ethics scholarship and provide a taxonomy of the field in terms of three overlapping categories: computational, philosophical, and critical.

[1] https://nips.cc/Conferences/2020/CallForPapers

- Carry an extensive conceptual analysis, and argue that critical approaches to AI ethics hold the most fruitful and promising way forward in terms of benefiting the most disproportionately impacted groups at the margins of society.
- Examine research publications from the two most prominent AI Ethics venues — FAccT and AIES — and present emerging trends and themes.

The rest of the paper is structured as follows. Section 2 provides a historical analysis outlining the existential tension between ethical perspectives that put emphasis on theoretical and abstract conceptualizations on the one hand, and those that work from concrete particulars on the other; highlighting the limitations of abstraction in the context of AI ethics. Section 2.1 outlines how critical approaches that interrogate systemic, societal and historical oppression, challenge existing power asymmetries and explicitly name the failures, negative societal and ethical implications of AI systems in a given application, group or context are likely to bring about actionable change. Section 3 presents the methodological details for our systematic analysis of the trends of the field through publications in the FAccT and AIES conferences. This is followed by a quantitative summary of the the main patterns of current publications in the two conferences' proceedings in Section 4. Section 5 takes an in-depth qualitative look at how one of the canonical works — the COMPAS algorithms — has been covered in the literature. Section 6 concludes the paper.

2 THE IRREMOVABLE TENSION BETWEEN THE ABSTRACT AND CONCRETE

The topic of ethics, particularly within the Western context — like other branches of philosophy — is often a speculative (rather than a practical) endeavour [28]. As a branch of Western philosophy, the broad field of ethics enquires, assesses, and theorizes about questions of justice, virtue, fairness, good and evil, and right and wrong. Such examinations often treat questions of ethics in an abstract manner, aspiring to arrive at universal theories or principles that can be uniformly applied regardless of time, place, or context [19, 47, 68]. Canonical Western approaches to ethics, from deontology to consequentialism, at their core strive for such universal and generalisable theories and principles "uncontaminated" by a particular culture, history, or context. Underlying this aspiration is the assumption that theories and principles of ethics can actually be disentangled from contingencies and abstracted in some form devoid of context, time, and space [18, 37]. What's more, this ambition for universal principles in ethics has been dominated by a particular mode of being human [2, 83]; one that takes a straight, white ontology as a foundation and recognizes the privileged Western white male as the standard, quintessential representative of humanity [2].

Traditions like Black Feminism [14, 34], Care Ethics [56], and the broad traditions of ethics that have emerged from Asian [4, 45] and African [49, 52, 71] scholars challenge the ontologies presumed in the "canonical" approach to ethics common to both the field of philosophy and the growing field of AI Ethics. These traditions are often grounded in down-to-earth problems and often strive to challenge underlying oppressive social structures and uneven power dynamics. Although there have been notable attempts to incorporate non-Western and feminist care ethics into the existing ethical paradigms within AI Ethics and Western philosophy [54, 74], these attempts remain marginalized by the structure of both disciplines. This situation is further exacerbated by the paucity of available expertise at the instructor level for less commonly taught philosophies, which has led to a dangerous reification of how philosophy, and ethics by extension, should be applied [75]. Given the ways that existing AI Ethics literature builds on the circulation of existing philosophical inquiry into ethics, the reproduction of the exclusion of marginalized philosophies and systems of ethics in AI Ethics is unsurprising, as is the maintenance of the white, Western ontologies upon which they are based.

Abstraction can be, and in some domains such as mathematics and art is, a vital process to discern and highlight important features, patterns, or aspects of a subject of enquiry — a way of seeing the forest for the trees. However, while abstractions and aspirations for universal principles can be worthy endeavours, whether such an objective is useful or suitable when it comes to thinking about ethics within the realm of AI and ML, is questionable. Scholars such as Heinz von Foerster [77] have argued that: *"Ethics is a matter of practice, of down-to-earth problems and not a matter of those categories and taxonomies that serve to fascinate the academic clubs and their specialists."* Selbst et al. [67] further contend that, when used to define notions of fairness and to produce fairness-aware learning algorithms, abstraction renders them "*ineffective, inaccurate, and sometimes dangerously misguided.*" Accordingly, the authors emphasize that abstraction, which is a fundamental concept of computing, also aids to sever Machine Learning tools from the social contexts in which they are embedded. The authors identify five traps, which they collectively name *the abstraction traps*, within the AI Ethics literature and the field of Computer Science in general, which result from failure to consider specific social contexts. They further note that: "*To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore to make a category error [...] an abstraction error.*" On a similar note, Crawford contends that abstraction mainly serves the already powerful in society *"The structures of power at the intersection of technology, capital and governance are well served by this narrow, abstracted analysis."* [16]

When algorithmic systems fail — and they often do — such failures often result in concrete negative outcomes such as job loss or exclusion from opportunities.Landmark work surrounding algorithmic decision-making in recent years has brought to light issues of discrimination, bias, and harm, demonstrating harms and injustices in a manner that is grounded in concrete practices, specific individuals, and groups. Computer Vision systems deployed in the social world suffer from disproportionate inaccuracies when it comes to recognizing faces of Black women compared to that of white men [11]; when social welfare decision-making processes are automated, the most vulnerable in society pay the heaviest price [23]; and search engines continually perpetuate negative stereotypes about women of colour [55]; to mention but a few examples. In short, when problems that arise from algorithmic decision-making are brought forth, we find that those disproportionately harmed or discriminated against are individuals and communities already at the margins of society.

continue from here–AB

This relation between AI (and AI-adjacent) systems and negative societal impact is not a new phenomenon. Indeed, one of the earlier police computing systems in the USA was used starting in the 1960s by the Kansas City police department to (among other tasks) analyze and allocate resources for operations and planning, which included police officer deployment [50]. By using not only attributes such as the race of individuals arrested, but also geographically linked data (such as number of crimes within a given area), the system created an increased surveillance and policing state that was particularly devastating for the Black population of Kansas City (see [65], and [73], for historical discussions of the laws and practices that led to geographic segregation by race that continues to be a feature of USA neighborhoods). This is not a particularly surprising outcome if one looks beyond strictly computing systems to consider how various technologies have been used across space and time to surveil and police Black people in the USA [9]. Similarly, Weizenbaum [80] argued in 1976 that the computer has from the beginning been a fundamentally conservative force which solidified existing power — in place of fundamental social changes. He argued that the computer renders technical solutions that allow existing power hierarchies to remain intact.

Tracking the relation of current AI systems to historical systems and technologies (and, indeed, social systems) allows us to more correctly contextualize AI systems (and by extension considerations of AI Ethics) within the existing milieu of socio-technical systems. This gives us an opportunity to account for the fact that *"social norms, ideologies, and*

practices are a constitutive part of technical design" [6, p.41]. Benjamin [6] makes a useful connection to the "Jim Crow" era in the United States in coining the term the "New Jim Code". She argues that we might consider race and racialization as, itself, a technology to *"sort, organize, and design a social structure*" [6, p.91]. Thus, AI Ethics work addressing fairness, bias, and discrimination must consider how a drive for efficiency, and even inclusion, might not actually move towards a "social good", but instead accelerate and strengthen technologies that further entrench conservative social structures and institutions. In approaching questions of AI ethics, social impact of AI and justice, examining the broader social, cultural, and economic context as well as explicit emphasis on concrete and potential harms is crucial. This focus on broader social factors and explicit emphasis on harms allows for the field to move towards bringing about material change and shifting power from the most to the least powerful in society, a vision Kalluri outlined in [38].

2.1 The limits of fairness

One of the ways ethics is studied in the AI space is through fairness. Fairness in algorithmic fairness is a slippery notion that is understood in many different ways, often with mutually incompatible conceptions [26]. While attempts to formalise the concept mathematically and operationalise quantitative measures of fairness may reduce the vagueness, such approaches necessarily treat fairness in abstract terms. The treatment of concepts such as fairness and bias in abstract terms is linked to the notion of neutrality and objectivity, the idea that there exists a "purely objective" dataset or a "neutral" representation of people, groups, and the social world independent of the observer/modeller [6, 7]. This is often subsequently followed by simplistic, reductive, and shallow efforts to "fix" such problems such as "debiasing" datasets — a practice that may result in more harm as it gives the illusion of a solution while the problem is rooted further deep [32]. Among other issues, such ways of thinking place problems on datasets and away from contingent wider societal and structural factors.

The discussion of "fairness" as an abstract mathematical or philosophical concept makes actual consequences from failures of AI systems a distant statistical or conceptual problems, effectively allowing us to overlook impacted individuals behind each data point [20, 63]. This is where analysis of power asymmetries proves critical: while fairness in abstract terms as a statistic or a figure on a table might be something the most privileged in society who are not impacted by AI systems prefer to do, for those who carry the burn of algorithmic failures, it is inseparable from their daily lives.

Algorithmic systems do indeed yield numerous benefits and hold the potential to drive positive social changes. However, without the acknowledgement that such systems tend to maintain conservative social structures [80] and produce uneven distribution of benefits and harms, blanket statements such as "beneficial AI" remain vacuous. This means, any discussion of the benefits and harms of AI need to be clear about who is benefiting and who is being harmed. Abstract statements such as "unfairness" or "bias" can serve as a smoke screen that allow specific harms such as racism, sexism, ableism, and other concrete harms to go unnamed.

An approach to ethics that ignores historical, societal, and structural issues omits factors of critical importance. In *"Don't ask if artificial intelligence is good or fair, ask how it shifts power"*, Kalluri [38] argues that genuine efforts to ethical thinking are those that bring about concrete material change to the conditions of the most marginalized, those approaches that shift power from the most to the least powerful in society. Without critical considerations of structural systems of power and oppression serving a central role in the ethical considerations of AI systems, researchers, designers, developers, and policy-makers risk equating processes and outcomes that are *fair* with those that are *equitable* and *just*. Understanding and accounting for these systems is needed to engage with the dangers of scale that AI and related algorithmic systems pose. This is especially true as AI systems themselves are also adapted to work within said

ongoing systems of power and oppression to efficiently scale the effects of those systems on marginalized populations (e.g., [23, 55]). Giving precedence to these critical considerations of structural systems and making them central to the considerations of the entire lifecycle of AI systems is required for the creation of equitable and just systems.

Despite the recent calls for a resistance to try to de-couple socio-cultural considerations from the (socio-)technical AI systems built and to sustain a critical perspective that take systems of power into consideration [17, 33, 67], the lack of such critical perspectives persists. Considerations of ethics and fairness in AI systems must make structural systems of power and oppression central to the design and development of such systems and resist the overly (techno)optimistic perspectives that assume context-less fairness will result in equitable outcomes. Furthermore, these considerations must directly contend with the possibility that AI systems may be inequitable and oppressive by their very application within particular contexts. The next section presents the current state of the field in relation to such critical analysis through the study of papers presented at two premiere AI Ethics conferences.

3 SURVEYING THE FIELD

Our approach "follows" AI Ethics through publications from two of the most prominent AI ethics venues; FAccT and AIES. Over the past five years, FAccT and AIES have risen to be the top two conferences where the community gathers to discuss, debate and publish state-of-the-art findings and ideas. By examining the publications that make up recent scholarship on AI Ethics, our review reveals the shape and contours of how "ethics" is conceptualized, operationalised, and researched and where it might be heading. Through this approach, we can come to understand the "work" that ethics does in the field of AI Ethics, where ethics "goes" insofar as how specific culturally-coded modes of understanding are circulated throughout the field, as well as the boundaries of what is and is not considered ethical. Following ethics in this way, therefore, provides crucial insights into the development of AI Ethics, and what the subject of legitimate AI Ethics inquiry should be.

3.1 Method

We read and annotated all papers published in the FAccT (formerly known as FAT*) and AIES conference proceedings over the past four years, from the founding of the conferences in 2018 to date (2021). This resulted in a total of 535 papers. Of those, we analysed 526 papers. For consistency, 91% of papers were annotated by two independent annotators. Overall, the annotators strongly agreed on all three of the measures. This is reflected by the Cohen's Kappa [13], a common inter-rater agreement measure, values of 0.93, 0.91, and 0.92 for the concreteness, disparity, and explicitness measures respectively. We recorded the paper title, authors, year of publication, publication venue, and any author-defined keywords. For each paper, we then read and annotated the abstract, introduction, and conclusion and/or discussion section according to an annotation scheme we developed (described below).

3.2 Annotation Scheme

Through an iterative process, we created an annotation scheme to classify and score each paper on three dimensions. We collectively established a set of guidelines that served as a point of reference for annotators during the annotation process. Each paper was annotated by two independent annotators each of whom completed a three-part annotation classification and scoring process as follows:

(1) **Concreteness Measure.** Annotate the paper along a binary concreteness measure with values *concrete* or *abstract*.

- A paper is labelled *abstract* if it approaches concepts such as ethics, fairness, harm, risk, and bias (shortened as ethics from here on) in a manner where context such as existing technological applications and/or discussion of specific impacted groups, classes, or demographics are absent. This might include presupposition of universal principles, invoking hypotheticals such as thought experiments, suppositions, if-then statements, and/or framing of such concepts as a primarily mathematical formulation when addressing the aforementioned ethical concepts (see Section 2 for further discussion on abstract and concrete).
- *Concrete*, on the other hand, is a label applied to a paper that approaches ethical concepts in a manner that is grounded in actual applications of a certain tool or technology, is rooted in specific demographics, living people, and/or culture and context-specific socio-technical phenomena.

(2) **Disparity Measure.** We used disparity measures to assess whether a paper investigates, addresses, and/or mentions the disparate impacts of an algorithmic system. We identified five dimensions defined as follows. Each paper was assigned one of the five values.
- *Strong* — the primary focus of the paper centers around exposing and tackling (through theoretical analysis, empirical experiments, and/or critical analysis) disparate impacts, possibly towards particular groups, classes, or demographics.
- *Medium* — the paper engages with disparate impacts (e.g., by addressing "protected attributes") in a substantial manner but falls short of making such discussion or analysis its primary objective.
- *Weak* — the paper mentions, perhaps in passing, disparate impact of algorithmic systems but lacks any significant engagement with the topic.
- *Agnostic* — the paper does not mention any impact of algorithmic systems on individuals, groups, communities, or demographics.
- *Inadequate* (*naive*, *ill-considered*, *ill-advised*, or *anti-equity*) — the paper may mention or even engage with the topic of disparate impact but does so in a way that shows a certain naivety of the topic or in the impact of the application of the algorithmic system, or the work in the paper may even explicitly enact disparate impacts, whether at the individual or institutional level. As the annotation scheme was developed in an iterative process, this dimension was added to the scheme at a later stage following a handful of recurring themes that emerged in the literature. Papers in this category also include those dealing with sensitive issues, such as work proposing the use of algorithms for identifying gang-related violent crimes with no discussion of social impact or harm the work might cause to individuals or groups deemed to be "gang memmbers".

(3) **Explicitness Measure.** For papers labelled with a disparity measure of *strong*, *medium* or *weak*, we further annotated that paper along a binary explicitness measure with the values *explicit* or *implicit*.
- A paper is labelled *explicit* if it unambiguously names the mentioned or referenced groups, communities, or demographics that are disproportionately impacted. The annotator then records the mentioned groups, communities, or demographics.
- The paper is labelled *implicit* if it discusses or mentions harm or impact without naming specific impacted groups, communities, or demographics and instead deals with harm in an implied, inferred, or indirect manner.

Some of the papers annotated were difficult to classify within our annotation scheme. For example, one paper focused on the social and ethical concerns of the Chinese Social Credit System [22]. Although this paper fits within the broad category of addressing the ethics of algorithmic systems and their societal implications, the framing, perspective, and central subject did not fit the annotation scheme we had outlined well. The paper discussed the Chinese Social Credit

system, what constitutes "good" and "bad", as well as what this means for the society. Although this was a compelling work, it was challenging to translate it into a traditional "fairness" template most papers within FAccT and AIES seem to adhere to. This is not to say that this (and the other papers) are outside the merits of AI Ethics, but rather a reflection of the limitation of our annotation scheme.

4 QUANTITATIVE SUMMARY

This section describes and discusses the results of our survey of papers from FAccT and AIES across the past 4 years (2018-2021). Sub-Section 4.1 shows the results and how the papers were distributed along the different dimensions we surveyed, and Sub-Section 4.2 provides an in-depth discussion of the results with relation to the larger context of the paper.

4.1 Results

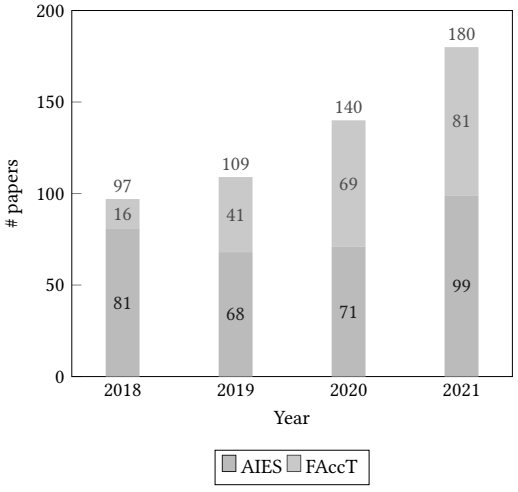

Fig. 1. Number of papers annotated by conference for each year.

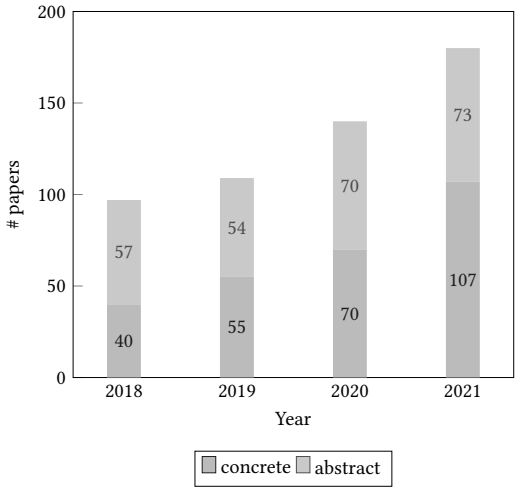

Fig. 2. Number of papers per concreteness value per year.

Figure 1 shows how the annotated papers are distributed across the 2 conferences and the 4 years considered. Each year, more papers relevant to our survey were published across the two conferences. While the number of papers from AIES saw some variations, the number of papers from FAccT grew significantly over the years.

Table 1. Percentage and number of *concrete* papers per conference per year.

Conference	2018		2019		2020		2021		Total	
	%	#	%	#	%	#	%	#	%	#
AIES	38	31	51	35	32	23	64	63	47.6	152
FAccT	56	9	49	20	68	47	54	44	58	120
Total	41	40	50	55	50	70	59	107	52	272

Figure 2 and Table 1 present results related to the concreteness measure. Figure 2 shows, for each year, the number of *concrete* and *abstract* papers. Table 1 details the distribution for each of the 2 conferences, showing both the absolute number and proportion of *concrete* papers in each conference for each year. The proportion of concrete papers grew slightly over the 4 considered years, from 41% in 2018 to 59% in 2021. Over this period, FAccT accepted proportionally more *concrete* papers than AIES.

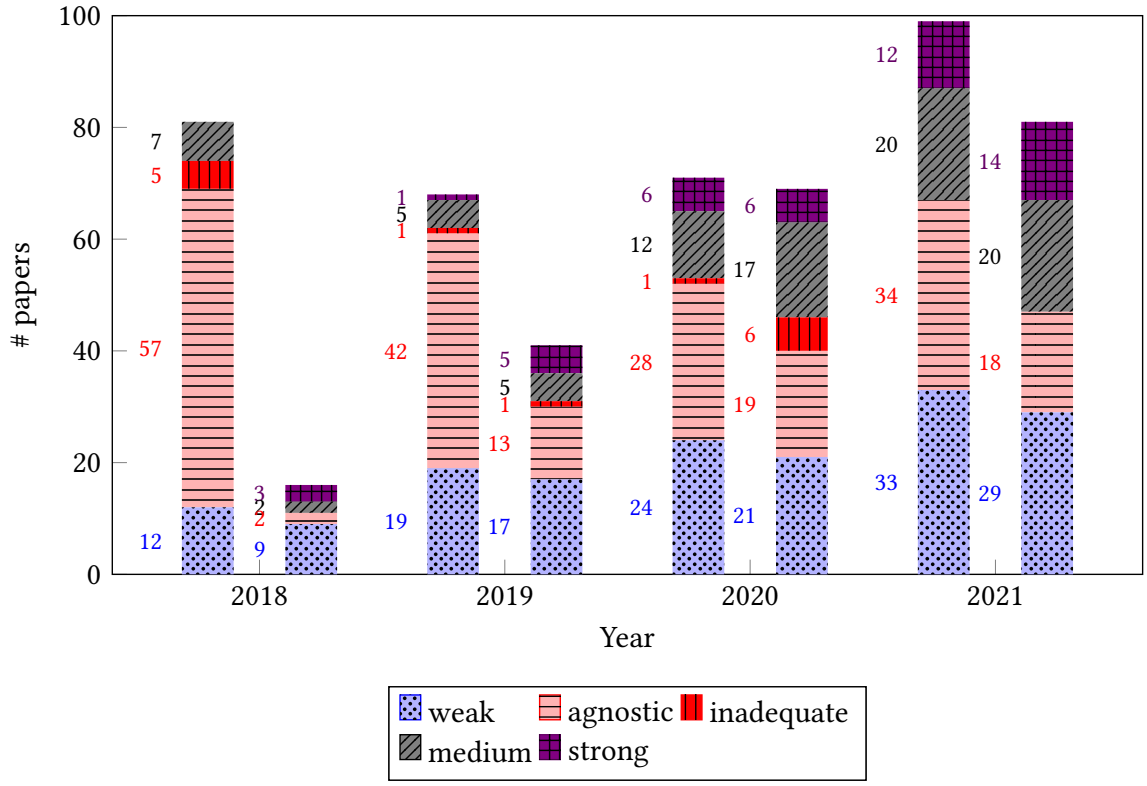

Fig. 3. Number of papers for each disparity measure value per year for AIES (left), and FAccT (right).

Table 2. Percentage and number of papers with agnostic position on disparate impact per conference and per year.

Conference	2018		2019		2020		2021		Total	
	%	#	%	#	%	#	%	#	%	#
AIES	70	57	61	42	38	27	32	32	49	158
FAccT	6	1	32	13	28	19	22	18	25	51

Figure 3 and Table 2 present results related to the disparity measure. Figure 3 shows the number of papers for each disparity measure value in each conference for each year. Overall, the *agnostic*, *weak*, and *medium* categories represent most of the papers, with 40%, 31%, and 17% of all papers respectively. This is particularly true in the earlier years. It

appears we may be seeing the beginning of a trend towards increased numbers of *medium* and *strong* papers in both venues as the proportion of *agnostic* papers that do not mention any disparate impact of algorithmic systems declines. Indeed, in 2018, the *medium* and *strong* categories represented 12% of papers, and the *agnostic* papers 61%; while in 2021, *medium* and *strong* covered 37% of papers and *agnostic* decreased to 29%. Table 2 focuses on the *agnostic* category and shows, for each conference and each year the number and proportion of these papers. The results show that for AIES a significant proportion of papers (49% overall) fall into this category, but that this proportion has decreased through the years.

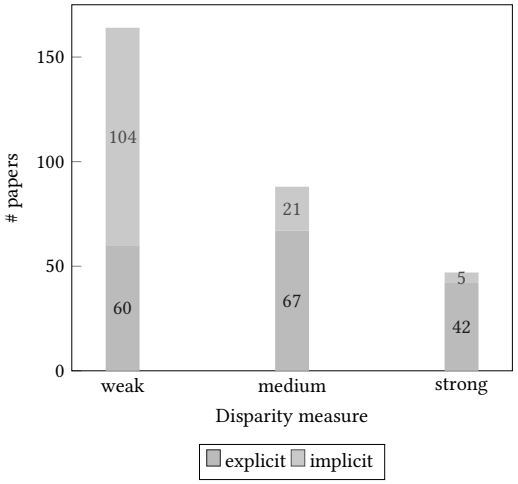

Fig. 4. Number of paper per explicitness per disparity measure value.

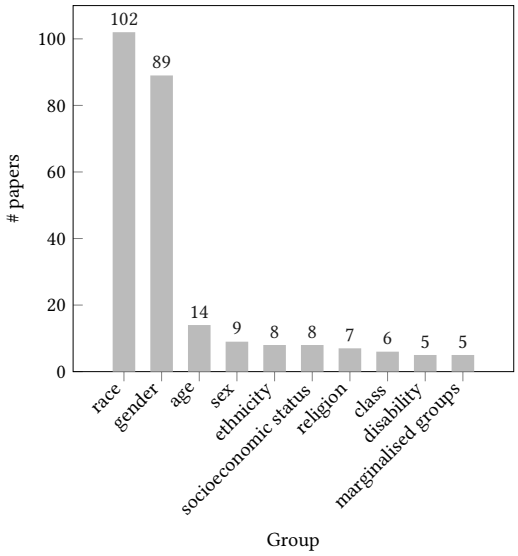

Fig. 5. 10 most mentioned impacted groups.

Figure 4 shows the breakdown of the explicitness measure across the three disparity categories to which it applies (*weak*, *medium*, and *strong*). Results show that most papers (62%) in the *weak* category use *implicit* disparity measures, while those in the *medium* and *strong* category tend to be more *explicit* in their discussions of algorithmic harm (76% and 89% respectively), unambiguously naming the communities and demographics that are the focus of study. Of those, Figure 5 shows the 10 groups that were most mentioned or discussed in the annotated papers, and for each group the number of papers that mentioned it. Race and gender are, by far, the two most mentioned impacted groups.

4.2 Discussion of the results

The results of this survey give insight into the current state and potential directions of AI Ethics scholarship at each conference. FAccT began as a much smaller venue, with a particular acknowledgement of algorithmic systems as socio-technical systems, as well as an early pointer to interdisciplinary work. This focus was reflected in the papers in the inaugural conference (see Table 1) where the majority of the (albeit small number of) papers were *concrete*. Furthermore, very few of those papers were *agnostic* and thus very few did not engage in some way with how the socio-technical system in focus related to some form of oppression or marginalization. Papers published in FAccT showed consistency across the years in (at least) naming particular groups that have been negatively impacted by socio-technical systems as can be seen from the non-*agnostic* counts in Fig. 3.

Though AIES has always had a multidisciplinary focus, most of the papers presented at the inaugural conference had a much more abstract focus, and the majority of those papers had an agnostic position. This abstract and agnostic focus perhaps points to less involvement from those disciplines that might center the experiences of marginalized communities when interacting with AI systems when the conference began. Many of the papers failed to engage with any intersections between the chosen topic within AI ethics in a way that was concrete or in a way that acknowledged the potential intersections between their system or approach and some form of oppression or marginalization. Furthermore, a relatively large portion of the papers during that inaugural conference were annotated as "inadequate" in some way, telling a troubling story of the initial wave of papers approaching this topic. Despite the troubling beginning to the proportion of inadequate papers, we did see an improvement, after the inaugural conference. The more recent shift towards a smaller proportion of *agnostic* papers and a higher proportion of both *strong* and *medium* papers hopefully points to an increasing recognition among that community of the importance of work that considers the interactions between the systems they discuss and build, marginalized communities, and the systems of oppression that result in this marginalized status.

Overall, the conferences appear to have more similar proceedings in terms of concreteness (Table 1) and non-*agnostic* (Table 2) positions in the most recent year (2021), than when they began. Though this does mark an improvement in some ways, the continued shallow engagement with systems and structures that define and result in marginalization is, nonetheless, concerning and problematic. That most papers consistently were either *agnostic* or *weak* (Fig. 3) indicates that much of the work is lacking in historical and ongoing contextualization of the algorithmic socio-technical systems being approached by those papers. We also see that papers with *strong* disparity measure tend to explicitly reference impacted groups the most (89%). This tendency decreased with *medium* papers (76%), while only a minority of *weak* papers were *explicit* (37%) in their naming of particular groups harmed or disparately impacted during their discussions (Fig. 4). This is a telling even if somewhat expected result — it is difficult for papers to not explicitly name impacted groups where the central aim of the paper centers around exposing or tackling disparate impacts. With *weak* papers, it is easier, and perhaps more likely, to see examples of work approaching disparate impact that does not explicitly name who is impacted and how they are impacted. The inadequacy seen with a majority of papers we surveyed being *agnostic* or *weak*, and thus either implicitly addressing harms or not addressing them at all, reflects an overall inadequacy of AI Ethics concerning the engagement with perspectives and histories that are not a part of the standard, or the Human [83].

The number of papers that we annotated as inadequate is relatively low compared to the total number of papers appearing at these venues, however their presence itself warrants further discussion. When looking at papers annotated as a category that fell under inadequate, most of those papers appeared in 2018 and 2020 (with AIES having the only papers rated in such a way during 2018 and FAccT having the majority of papers falling in that category in 2020). This small subset of papers were varied in the ways in which they were inadequate—-some failed to consider potentially problematic uses of their systems given the context in which they discussed them, while others seemed to positively engage with perspectives that could be viewed as anti-equity. One example of the former is presented in Rodolfa et al. [64]. The authors present a system they developed in collaboration with the Los Angeles City's attorney office that uses predictive modeling to prepare social services resources for those likely to be re-arrested (with a focus on those with prior interactions and arrests with the police). Understandably, the authors want to find a better use of these predictive algorithms than re-arrest and the authors include a healthy discussion on equity in the predictive models. However, the considerations of how these predictive algorithms fit into the institution-person interaction, especially from the perspective of the person who has been re-arrested (or may be in the future) are extremely limited. At its core, this inadequacy stems from a shallow consideration of bias that does not more readily approach the racial (and

interlocking class and gender) systems that undergird an individuals interactions with these criminal justice institutions. Without such considerations, there is a risk of inadequate considerations of topics like how the predictive modeling system may be adapted within an existing institution that might be, for example, anti-Black.

Another paper that fell under the *inadequate* category discussed affirmative action. The paper by Kannan et al. [40] uses computational modeling to explore aspects of downstream effects of affirmative action policies in undergraduate college admissions. In the limitations section of the paper they note that their approach does not account for "past inequity" and "...does not attempt to address or correct the historical forces...". While the goal of finding a new way to understand effects of these policies is understandable, failing to consider these critical forces appears to amount to color-blind exploration, which is likely be used as way to bolster current systems of racism [8]. Thus, such assumptions may cause the work itself (and continuing work) to be antithetical to some of the movements that progressed policy such as affirmative action (e.g., see [42]); such *limitations* assume away the very structural systems that are a large part of the inequity.

5 QUALITATIVE ANALYSIS: PROPUBLICA'S COMPAS AS A CASE STUDY

The ProPublica COMPAS study [5] was one of the most recurring citations in the papers we studied (e.g., [11, 12, 25, 29, 31, 39, 41, 43, 48, 51, 53, 70, 72, 76]). We found that many of the papers failed to engage beyond a shallow mention of the paper or the associated data. Instead, these papers frequently focused on ways to improve fairness in assessing likelihood to re-offend or, more generally, ways to increase fairness within current incarceration processes, with authors finding ways within their specific approaches to train a system that produced fairer results.

However, those perspectives often failed to address what it means to re-offend within a society (particularly in the context of the US criminal justice system) that deems some people more likely to offend than others (e.g., due to their race, sex, gender, class status, and intersections thereof or, relatedly, due to their "governability" [62]). These mentions typically failed to question how datasets are sourced and curated [61], what was used as input data to build the algorithm, how historical and current stereotypes are encoded [6], and whether improving recidivism algorithms nonetheless results in a net harm to marginalized groups.

Taskesen et al. [72] can be used as one example for critique, whereby the critique is recontextualized to other papers that we analyzed. In this paper, Taskesen et al. [72] describe a framework for assessing probabilistic fairness for classifiers. The authors note that we have recently seen "data and algorithms" become "an integrative part of the human society" and use the COMPAS dataset to provide an exemplar of their proposed statistical test of fairness in use. While, commendably, the paper does focus on fairness and test it on the real-world dataset, it critically lacks in its contextualization of the very systems that help define the problem space and environment in which the classifiers (that they hope to assess for fairness) operate. This lack of context obscures a very important question of how probabilistic fairness interacts with the historical and ongoing dehumanizing processes; those processes that help define who is functionally deemed a part of "human society". The lack of a historical context also results in incorrectly labeling *the use of* datasets and algorithms as a new practice. Although, our use of digital systems to represent, extend, and modify data and algorithms has certainly changed, what we do with digital systems now is a continuation of decades old *computing* systems. Though the paper does conclude a note on problematic potential of auditing tools falling in the hands of "vicious" inspectors, such an individual focus fails to consider that such audits occur within a context of systems that impose hierarchies determined by hegemonic knowledge. That is, the use of the word "vicious" obscures the mundaneness of the violence enacted by these systems and thus the normalcy of an auditor's behavior when they use such a tool in a way that results in (racial, patriarchal, etc.) violence.

Staying in the context of the COMPAS dataset, it is less impactful to consider being fairer in the ratings of likelihood of re-offense if we fail to explicitly take into account why those who are Black are more likely to be regarded as a criminal, and this may be due to their physical or social proximity to other previous "offenders", or to spaces that have historically excluded them, for example racially and class segregated neighborhoods [2]. These perspectives fail to adequately address how the institutional contexts in which those systems are applied (i.e., as socio-technical systems) bring with them a socio-cultural milieu that defines not only how those systems are implemented and integrated, but also how those institutions themselves adapt to any given deployment or integration of technology. For example, they often fail to critically address the antiblackness that undergirds the socio-technical system they hope to improve (e.g., see [17] for a discussion of the need to move beyond notions of representation and bias in the development of AI systems). More importantly, the fixation on improving such algorithms by tweaking them to make them fairer or more accurate, contributes to maintaining punitory systems, rendering reformative and transformative questions such as "what mechanisms would be put in place to help the convicted integrate back into society after release?" out of the purview of AI ethics. These nuances, perspectives, and important backgrounds are lost as we approach ethics, fairness, and related concepts as solely abstract mathematical formulations or philosophical musings. An approach cannot adequately address this adaptability unless it engages with the socio-cultural institutions that will ultimately enact it.

Having said that, acknowledgement of systems of oppression in the context of COMPAS is not entirely absent. Fogliato et al [25] and Krafft et al. [43], for example, provide examples of more directly engaging with issues related to systems of oppression and ideas of power, even if they do so with different aims. Fogliato et al [25] present their investigation of biases (particularly racial) in arrest data that might be used by risk assessment instruments (such as COMPAS). Though the paper falls short in some areas of addressing antiblackness as a concept directly [3], they did discuss their work in the context of related concepts, such as victim devaluing and police response to criminal acts in racially segregated (predominantly Black) areas. Krafft et al. [43] describe the development and use of an AI auditing toolkit that can be used as a development and deployment decision tool. Deviating from typical practice, the design and development of this toolkit explicitly engaged with a wider audience, including community organizers and those with the lived experience of marginalization; this community orientation helps to center the issues faced by those communities. Though both of these papers have a different stated aim than many of the other papers we reviewed that discuss and use COMPAS and the Propublica study, we would argue that work in this space should have a goal of finding ways to center the participation from the communities affected by the systems they wish to make more "fair", and more of a focus on centering the historical and ongoing oppressive systems and institutions that result in disproportionate harm on those marginalized and dehumanized communities.

6 CONCLUSION — MISSING FOREST FOR THE TREES

Even though AI Ethics is a fast growing and broad field of enquiry, its pace is no match for the rate at which algorithmic systems are being developed and integrated into every possible corner of society. The broadly construed field of AI Ethics holds a crucial place to ensure development and deployment of algorithmic systems that are just, equitable, and beneficial to all members of society, including those most negatively impacted by inequality. A review of the AI ethics papers from two of the most prominent conferences showed there is a great tendency for abstract discussion of

[2] See [30] for a related discussion on the problem of *innocence*, as well as [79] for a discussion of relations between Blackness and Law.
[3] For example, in their analysis of the city of Memphis, the authors could have noted and traced historical antiblackness [81] and how surveillance [35] and violence enacted by policing systems relates.

such topics devoid of structural and social factors and specific and potential harms. Given that the most marginalized in society are the most impacted when algorithmic systems fail, we contend that all AI ethics work, from research, to policy, to governance, should pay attention to structural factors and actual existing harms. In this paper, we have argued that given oppressive social structures, power asymmetries, and the uneven harm and benefit distributions , work in AI Ethics should pay more attention to these broad social, historical, and structural factors in order to bring about actual change that benefits the least powerful in society.

ACKNOWLEDGMENTS

This work was supported, in part, by Science Foundation Ireland grant 13/RC/2094_P2, and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero - the Science Foundation Ireland Research Centre for Software (www.lero.ie). This work was also partially supported by the National Science Foundation under grant No. 2218226.

REFERENCES

[1] Philip E Agre. 1994. Surveillance and capture: Two models of privacy. *The Information Society* 10, 2 (1994), 101–127.
[2] Sara Ahmed. 2007. A phenomenology of whiteness. *Feminist theory* 8, 2 (2007), 149–168.
[3] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[4] Roger T. Ames and David L. Hall. 2003. *Dao De Jing—Making This Life Significant—A Philosophical Translation.* Ballantine Books.
[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.
[6] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code.* John Wiley & Sons.
[7] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), 100205.
[8] E. Bonilla-Silva. 2015. The Structure of Racism in Color-Blind, "Post-Racial" America. *American Behavioral Scientist* 59, 11 (2015), 1358–1376. https://doi.org/10.1177/0002764215586826
[9] Simone Browne. 2015. *Dark matters: On the surveillance of blackness.* Duke University Press.
[10] Matthew Browning and Bruce Arrigo. 2020. Stop and Risk: Policing, Data, and the Digital Age of Discrimination. *American Journal of Criminal Justice* (2020), 1–19.
[11] J. Buolamwini and T. Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press
[12] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, New York, NY, USA, 134–148. http://proceedings.mlr.press
[13] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.
[14] Patricia Hill Collins. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment.* routledge.
[15] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need.*
[16] Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.
[17] C. L. Dancy and P. K. Saucier. 2021. AI and Blackness: Towards moving beyond bias and representation. *IEEE Transactions on Technology and Society* (2021), 1–10. https://doi.org/10.1109/TTS.2021.3125998
[18] Lorraine Daston. 2018. Calculation and the division of labor, 1750-1950. *Bulletin of the German Historical Institute* 62 (2018), 9–30.
[19] Hanne De Jaegher. 2019. Loving and knowing: Reflections for an engaged epistemology. *Phenomenology and the Cognitive Sciences* (2019), 1–24.
[20] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.
[21] Hubert Dreyfus. 1972. What computers can't do: The limits of artificial intelligence. (1972).
[22] Severin Engelmann, Mo Chen, Felix Fischer, Ching-yu Kao, and Jens Grossklags. 2019. Clear Sanctions, Vague Rewards: How China's Social Credit System Currently Defines" Good" and" Bad" Behavior. In *Proceedings of the conference on fairness, accountability, and transparency.* 69–78.
[23] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.
[24] Casey Fiesler, Natalie Garrett, and Nathan Beard. 2020. What do We teach when We teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education.* 289–295.
[25] R. Fogliato, A. Xiang, Z. Lipton, D. Nagin, and A. Chouldechova. 2021. On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. In *2021 AAAI/ACM Conference on AI, Ethics, and Society.* Association for Computing Machinery, 100–111. https://doi.org/10.1145/3461702.3462538

[26] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[27] Oscar H Gandy Jr. 2021. *The panoptic sort: A political economy of personal information.* Oxford University Press.

[28] Michael Gardiner. 1996. Alterity and ethics: A dialogical perspective. *Theory, Culture & Society* 13, 2 (1996), 121–143.

[29] Thomas Krendl Gilbert and Yonatan Mintz. 2019. Epistemic Therapy for Bias in Automated Decision-Making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 61–67. https://doi.org/10.1145/3306618.3314294

[30] R. W. Gilmore. 2017. Abolition geography and the problem of innocence. In *Futures of Black Radicalism*, G. T. Johnson and A. Lubin (Eds.). Verso, Brooklyn, NY, 225–240.

[31] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-Discriminatory Machine Learning through Convex Fairness Criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 116. https://doi.org/10.1145/3278721.3278722

[32] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 609–614.

[33] L. M. Hampton. [n.d.]. Black Feminist Musings on Algorithmic Oppression. In *2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 1. https://doi.org/10.1145/3442188.3445929

[34] bell hooks. 1984. *Feminist Theory: From Margin to Center.* Routledge.

[35] B. Jefferson. 2020. Digitize and Punish: Racial criminalization in the digital age. University of Minnesota Press, Minneapolis, MN, USA, 171.

[36] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[37] Alicia Juarrero. 2000. Dynamics in action: Intentional behavior as a complex system. *Emergence* 2, 2 (2000), 24–57.

[38] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169–169.

[39] Shivaram Kalyanakrishnan, Rahul Alex Panicker, Sarayu Natarajan, and Shreya Rao. 2018. Opportunities and Challenges for Artificial Intelligence in India. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 164–170. https://doi.org/10.1145/3278721.3278738

[40] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 240–248. https://doi.org/10.1145/3287560.3287578

[41] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/3306618.3314287

[42] M. L. King Jr. 1968. *Where do we go from her: Chaos or community?* Beacon Press, Boston, MA, USA.

[43] P. M. Krafft, Meg Young, Michael Katell, Jennifer E. Lee, Shankar Narayan, Micah Epstein, Dharma Dailey, Bernease Herman, Aaron Tam, Vivian Guetler, Corinne Bintz, Daniella Raz, Pa Ousman Jobe, Franziska Putz, Brian Robick, and Bissan Barghouti. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 772–781. https://doi.org/10.1145/3442188.3445938

[44] Peter Kuhn and Kailing Shen. 2013. Gender discrimination in job ads: Evidence from china. *The Quarterly Journal of Economics* 128, 1 (2013), 287–336.

[45] Chenyang Li. 1994. The Confucian Concept of Jen and the Feminist Ethic of Care: A Comparative Study. *Hypatia: A Journal of Feminist Philosophy* 9, 1 (1994), 70–89.

[46] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[47] Ivana Marková. 2016. *The dialogical mind: Common sense and ethics.* Cambridge University Press.

[48] Jeanna Matthews, Marzieh Babaeianjelodar, Stephen Lorenz, Abigail Matthews, Mariama Njie, Nathaniel Adams, Dan Krane, Jessica Goldthwaite, and Clinton Hughes. 2019. The Right To Confront Your Accusers: Opening the Black Box of Forensic DNA Software. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 321–327. https://doi.org/10.1145/3306618.3314279

[49] John S Mbiti. 1969. African Religions and Philosophy. New York: Frederick A.

[50] Charlton D McIlwain. 2019. *Black software: The internet and racial justice, from the AfroNet to Black Lives Matter.* Oxford University Press, USA.

[51] Daniel McNamara. 2019. Equalized Odds Implies Partially Equalized Outcomes Under Realistic Assumptions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 313–320. https://doi.org/10.1145/3306618.3314290

[52] Sabelo Mhlambi. 2020. From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. *Carr Center for Human Rights Policy Discussion Paper Series* (2020).

[53] Alan Mishler. 2019. Modeling Risk and Achieving Algorithmic Fairness Using Potential Outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 555–556. https://doi.org/10.1145/3306618.3314323

[54] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684.

[55] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

[56] N Noddings. 2013. *Caring: A Relational Approach to Ethics and Moral Education.* University of California Press.

[57] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[58] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.

[59] Frank Pasquale. 2015. *The black box society.* Harvard University Press.

[60] Neil R Powe. 2020. Black kidney function matters: use or misuse of race? *Jama* 324, 8 (2020), 737–738.

[61] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923* (2020).

[62] H. L. T. Quan. 2017. "It's hard to stop rebels that time travel": Democratic living and the radical reimagining of old worlds. In *Futures of Black Radicalism*, G. T. Johnson and A. Lubin (Eds.). Verso, Brooklyn, NY, 173–193.

[63] Inioluwa Deborah Raji. 2020. The discomfort of death counts: Mourning through the distorted lens of reported COVID-19 death data. *Patterns* 1, 4 (2020), 100066.

[64] K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani. [n.d.]. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 142–153. https://doi.org/10.1145/3351095.3372863

[65] Richard Rothstein. 2017. *The color of law: A forgotten history of how our government segregated America.* Liveright Publishing.

[66] R Joshua Scannell. 2019. This is not Minority Report: Predictive policing and population racism. *Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life* (2019), 107–129.

[67] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA, 59–68.

[68] John Shotter. 2006. Vygotsky, Bakhtin, Goethe: Consciousness and the dynamics of voice. (2006).

[69] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, New York, NY, USA, 5–19. http://proceedings.mlr.press/v81

[70] Biplav Srivastava and Francesca Rossi. 2018. Towards Composable Bias Rating of AI Services. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 284–289. https://doi.org/10.1145/3278721.3278744

[71] Sylvia Tamale. 2020. *Decolonization and Afro-feminism.* Daraja Press.

[72] B. Taskesen, J. Blanchet, D. Kuhn, and V. A. Nguyen. 2021. A Statistical Test for Probabilistic Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 648–665. https://doi.org/10.1145/3442188.3445927

[73] Keeanga-Yamahtta Taylor. 2019. *Race for profit: How banks and the real estate industry undermined black homeownership.* UNC Press Books.

[74] Shannon Vallor. 2016. *Technology and the Virtues: A Philosophical Guide to a World Worth Wanting.* Oxford University Press.

[75] B Van Norden. 2017. *Taking Back Philosophy: A Multicultural Manifesto.* Columbia University Press.

[76] Marisa Vasconcelos, Carlos Cardonha, and Bernardo Gonçalves. 2018. Modeling Epistemological Principles for Bias Mitigation in AI Systems: An Illustration in Hiring Decisions. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 323–329. https://doi.org/10.1145/3278721.3278751

[77] Heinz Von Foerster and Bernhard Pörksen. 2002. *Understanding systems: Conversations on epistemology and ethics.* Kluwer Academic/Plenum Publishers New York.

[78] Darshali A Vyas, Leo G Eisenstein, and David S Jones. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms.

[79] Alexander G Weheliye. 2014. Law: Property. In *Habeas viscus: Racializing assemblages, biopolitics, and black feminist theories of the human*. Duke University Press, Durham, NC, 74–88.

[80] Joseph Weizenbaum. 1976. Computer power and human reason: From judgment to calculation. (1976).

[81] Ida B. Wells. 1999 (1893). *Lynch Law.* University of Illinois Press, Urbana, Book section 4.

[82] Terry Winograd and Fernando Flores. 1986. *Understanding computers and cognition: A new foundation for design.* Intellect Books.

[83] S. Wynter. 2003. Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation - An Argument. *CR: The New Centennial Review* 3, 3 (2003), 257–337. http://www.jstor.org/stable/41949874

[84] Shoshana Zuboff. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* Profile Books.