AI & SOCIETY (2021) 36:1047–1056
https://doi.org/10.1007/s00146-021-01153-9

OPEN FORUM

Conservative AI and social inequality: conceptualizing alternatives to bias through social theory

Mike Zajko[1]

Received: 14 July 2020 / Accepted: 6 January 2021 / Published online: 7 February 2021

Abstract
In response to calls for greater interdisciplinary involvement from the social sciences and humanities in the development, governance, and study of artificial intelligence systems, this paper presents one sociologist's view on the problem of algorithmic bias and the reproduction of societal bias. Discussions of bias in AI cover much of the same conceptual terrain that sociologists studying inequality have long understood using more specific terms and theories. Concerns over reproducing societal bias should be informed by an understanding of the ways that inequality is continually reproduced in society—processes that AI systems are either complicit in, or can be designed to disrupt and counter. The contrast presented here is between conservative and radical approaches to AI, with conservatism referring to dominant tendencies that reproduce and strengthen the status quo, while radical approaches work to disrupt systemic forms of inequality. The limitations of a conservative approach to racial bias are discussed through the specific example of biased criminal risk assessments and Indigenous overrepresentation in Canada's criminal justice system. This illustrates the dangers of treating racial bias as a generalizable problem and equality as a generalizable solution, emphasizing the importance of considering inequality in context. Societal issues can no longer be out of scope for AI and machine learning, given the impact of these systems on human lives. This requires engagement with a growing body of critical AI scholarship that goes beyond biased data to analyze structured ways of perpetuating inequality, opening up the possibility for interdisciplinary engagement and radical alternatives.

Keywords Bias · Inequality · Sociology · Fairness · Politics · Colonialism

1 Introduction

As the profound consequences of AI-related technologies have become more widely recognized, AI practitioners are increasingly expected to address the ethical or political consequences of their work (Johnson 2020a), raising questions that have traditionally been in the domain of the social sciences and humanities. Scholars have documented the ways that automated decisions are depriving people of government benefits, discriminating on the basis or sex, skin color, age and numerous other forms of difference, choosing who is surveilled, who is imprisoned, or who is targeted for economic exploitation (O'Neil 2016; Gillespie and Seaver 2016; Eubanks 2017). Systems created with the promise of unbiased, objective judgment end up reproducing the biases and inequalities of the societies they are 'trained' on (Whittaker et al. 2018; Hoffmann 2019). In contrast to debates over the ethics of "artificial minds" or how AI might operate as a moral agent (Nath and Sahu 2020), the most prominent ethical concerns over today's AI systems parallel fundamental critiques of technology and politics (Winner 1980; Cooley 1995), asking how AI systems relate to existing power structures.

Technologists are often poorly prepared for these considerations, and dominant paradigms in data science have been criticized as narrow technical approaches to social problems, necessitating involvement from additional perspectives (Green and Hu 2018). Social sciences and humanities scholars have been identified as having salient skills and insights to contribute to AI's holistic development (Hartley 2017; G7 Science Academies 2019), particularly in respect to ethics, fairness, and bias (Lepri et al. 2018; Silberg and Manyika 2019; Kusner and Loftus 2020). I argue that one way interdisciplinarity can benefit these discussions is by pushing beyond the conventional understanding of bias in

✉ Mike Zajko
mike.zajko@ubc.ca

[1] Department of History and Sociology, Okanagan Campus, The University of British Columbia, Kelowna, BC, Canada

AI so as to better articulate the social good and the obstacles to achieving it. If politics is primarily about power, then traditional approaches in computing and data science are politically conservative in that they affirm existing power relations. AI has the potential to disrupt various institutions and social processes, but is typically used as a tool to reinforce the status quo and benefit those at the center, rather than the margins.

In this article, I examine the problem of societal bias and alternative ways of understanding the problem—chiefly as one of social inequality. While some definitions of bias focus on the accuracy of prediction or categorization, societal bias indicates some undesirable state of affairs, but without a basis for imagining what is desirable. Engaging with literature on specific forms of social inequality can lead to more productive conceptualizations of such problems, helping us to imagine futures that are not limited to the removal of bias. This can involve locating AI systems in relation to pre-existing social structures, and asking how those structures can be reinforced, reformed, dismantled or replaced.

Inequality has a history, and a sociological analysis can help us understand the sources of un-fairness in society (Rosanvallon 2013; Machin and Stehr 2016). Structural changes to inequality are usually best pursued by political means that do not depend on technological design, but for those who are approaching these problems as a design challenge, inequality needs to be considered first, rather than as an afterthought (when data or decisions are revealed as biased). The least that AI researchers and developers can do is to learn more about the human categories they are working with from fields that have a well-developed understanding of phenomena such as gender, race, class, or criminality, and how these are distributed and reproduced in society. This allows us to translate concerns such as 'gender bias' or 'racial bias' into concerns about specific processes and structures that AI systems are involved in.

In this article, I use the example of settler colonialism as one such social structure. Colonialism can be addressed through fairer algorithms or more equal decisions within existing processes, but these are still likely to perpetuate existing inequalities and to have largely conservative effects. Because society is structured along some fundamentally unequal lines, changes are more effective 'upstream' of existing AI systems, where new technologies are often not required. While we should be skeptical of attempts to solve the problem of inequality through new technologies, there may be ways of designing new AI systems that help to shift power, as long as this is done with the participation of the people, groups, and communities that such efforts are intended to help. In the example of racial inequality in a colonial context, using social theory to conceptualize the problem leads us to a very different starting place than the problem of bias and its removal.

2 When systems are biased

This section explains the problem of bias in AI, both in terms of how it is commonly presented in the field, as well as interdisciplinary alternatives. Complicating the first part of this task is the fact that AI researchers do not agree on what bias is or what to do about it, and many authors do not even attempt to define the term. In literature on data science, machine learning (ML), and AI, the implicit general definition of bias is any tendency, pattern or association that is problematic. There are examples of texts that discuss "useful bias", or biases that are desirable, particularly when discussing some human predispositions (Shah et al. 2019b), but generally when bias is invoked, it is a problem.

When authors specify forms of bias, these are typically differentiated by their source: human bias, machine bias, systemic bias, societal bias, historical bias, sampling bias, observation bias, and so on (for example, Shah et al. 2019a). There is considerable conceptual confusion and overlap with these terms, but they are used to distinguish, where the bias is supposedly 'coming from'; whether an individual decision maker's unconscious bias, historically entrenched distributions, or the issues with data collection and measurement. Whatever the source, all relate to at least one fundamental definition of bias, as being either (1) inaccurate, or (2) undesirable. This is roughly in line with Mitchell et al.'s (2020) distinction between "statistical bias" and "societal bias"—the former being a "mismatch between the sample used to train a predictive model, and the world as it currently is", while the latter relates to "concerns about objectionable social structures" (p. 4).

The most straightforward conceptions of bias to operationalize are statistical in nature, and based on the accuracy of representations or predictions – i.e., does the data accurately reflect the 'ground truth'? Does the model accurately predict the risk profiles or recidivism rates for different populations? Is the person with the best qualifications for the position accurately identified, or does some irrelevant factor bias the decision against them? These are questions that can be addressed with better data and models, overcoming problems with sampling, measurement, and design. However, the vast amorphous terrain of societal bias includes inequalities and injustices that can indeed be accurately reproduced, and therefore reinforced, by an algorithm. These include cases, where the "training data is tainted with historical bias" or where there is an "unequal ground truth" (Hacker 2018, p. 1148).

Gender bias exists even where an algorithm accurately reproduces cultural assumptions about gender embedded in language, returns photos of men for an image search of an occupation that is male-dominated, or predicts that

an individual is a man or woman on the basis of their occupation (see Suresh and Guttag 2020). It is sometimes possible to reframe these issues in terms of accuracy—perhaps the algorithm does not accurately reflect some fundamental equality between men and women—but when authors are concerned about algorithms that 'reproduce bias', this can include the accurate reproduction of existing social patterns and distributions. In these conceptions, bias is taken to mean whatever is "unfair", "undesirable", or "unwanted", which typically manifests as "systematic discrimination… based on the inappropriate use of certain traits or characteristics" (Silberg and Manyika 2019, p. 2). The easiest way to operationalize this second definition of bias is to tie it to some conception of fairness—hence any tendency we can consider unfair is also biased (Friedman and Nissenbaum 1996, p. 343). However, this does not get us very far because of multiple contradictory conceptions of fairness (Friedler et al. 2016).

Defining bias in terms of (1) statistical accuracy of prediction/estimation is well-suited for technical problem-solving, but not for addressing the widespread concern over 'reproducing biases' in society (Hoffmann 2019). The most inclusive (2) definition of bias is a tendency that is undesirable or unwanted, based on a "normative concern with the state of the world" (Suresh and Guttag 2020, p. 2), but this tells us nothing about what is desirable. Addressing this kind of bias involves either selecting one of the formal definitions of fairness, or some ad-hoc notion of a desirable outcome. There is no consensus among AI developers over what formal definition of fairness is best, and minimizing bias through some formal procedure can actually introduce new kinds of harm against individuals and groups—as has been found in jurisdictions attempting to treat men and women equally in predicting recidivism (see Corbett-Davies and Goel 2018). Desirable outcomes that are not based on formalizing fairness often rest on vague notions of 'doing good', but to move beyond the status quo we need to actually envision and debate the kind of world that we want to create (Green 2019). The language of bias is inadequate for this task. Even where 'debiasing' produces positive outcomes, this is not equivalent to social justice, democratization, decolonization, or other well-articulated political objectives. The language of bias does not help us attend to how capitalism, patriarchy, white supremacy, or colonialism are organized, and makes this work more difficult. Bias places the focus on circumstances of disadvantage, rather than "the normative conditions that produce—and promote the qualities or interests of—advantaged subjects" (Hoffmann 2019, p. 907). In other words, we end up focusing on systematically disadvantaged groups, rather than the reasons why these disadvantages exist (to produce systematic advantages for others), and this leads to conservative outcomes.

2.1 Conservative bias

AI systems are conservative by default in their political effects, although in data science this is widely understood as an issue of 'reproducing bias'. Because it acts as an obstacle to deeper, more structural analyses, using the term 'bias' actually reinforces this conservative tendency—by which I mean the conservation and maintenance of the social (political and economic) order. This basic definition of conservatism should not be equated with the various values that have historically been associated with conservative thinkers and political movements—a history that includes calls for social change as well as conservation (see Bourke 2018). According to conservative sociologist Robert Nisbet (1952), conservatism developed to emphasize the primacy of society or the social group as a response to the French Revolution, and in opposition to individualism, liberalism, and radicalism. In this article, I use conservatism to refer to an orientation that maintains the status quo in society, where radicalism is the opposing possibility of transformation or profound change (see Wolfe 1923). Conservatism is not inherently bad or harmful; there are aspects of the existing order that are worth keeping or reinforcing, but there are also many things about society that should be changed or improved. The problem lies in the limitations that a conservative approach imposes on social change and the possibilities it forecloses.

Work that addresses bias in AI has a tendency to "optimize the status quo" (Carr 2014), promoting changes that preserve existing inequalities. Abeba Birhane (2020) has been one of very few to actually name this tendency as being 'conservative', but there is widespread recognition that AI, as it currently exists, reproduces existing social distinctions and biases. There are multiple reasons for these conservative tendencies, including the fact that AI development tends to follow the logic of capitalism (Chiang 2017), leading to systems that 'think' as a corporation would (Penn 2018). Institutionalized cultures in academia have also often reinforced a narrow focus (for instance, through formalism in computer science, see Leith 1990; Selbst et al. 2019; Green and Viljoen 2020), and academic projects must be carefully aligned with sources of funding, often including close ties with private industry (Hoffman 2017). Computing and data science programs often have the same issues with diversity as private industry, so that many technologies are developed by teams from similarly privileged backgrounds. This results in a "feedback loop" (West et al. 2019) through the design of conservative AI systems that reproduce unequal and discriminatory effects. All of this is exacerbated by attempts to remain politically neutral and objective, where "objectivity and neutrality do not mean value-free—they instead mean acquiescence to dominant scientific, social, and political values" (Green and Viljoen 2020, p. 22).

There are practical and conceptual ways to go beyond conservative approaches, which involve finding other ways of talking and thinking about bias. This means that societal bias needs to be brought 'into scope' for AI development, and doing so requires engaging with theories developed through decades of work on these issues, providing us with a richer conceptual vocabulary to specify problems and imagine better futures. In social theory, bias is rarely discussed, and when the term does appear, it is in the context of the reproduction of inequality and the status quo. For example, Steven Lukes, in *Power: A Radical View*, discusses "organization [as] the mobilization of bias" (Bachrach and Baratz; Schattschneider, as cited in Lukes 2005, p. 20), to characterize how power works through institutionalized procedures that favor some groups over others. Rather than referring to inaccuracies or flaws in decision-making, this conceptualization of bias relates to how institutions perpetuate social hierarchies (Goetz 1997). It is this broader view of inequality that social theory can inform, for which the specific concept of bias is often quite dispensable.

3 Interdisciplinary alternatives to the problem of bias

In what follows, I offer an interdisciplinary contribution to the topic of societal bias in AI, where interdisciplinarity is used to formulate new problems (Lury 2018), rather than providing new solutions to existing problems. Even in situations, where two disciplines use different language to discuss what might appear to be the same problem, switching from one disciplinary discourse to another can significantly shift how problems are formulated. This is certainly the case when it comes to questions of bias and fairness in AI, where the language of 'societal bias' and 'unequal ground truth' could benefit from being transformed or replaced by more elaborated concepts in social theory related to inequality. To move beyond some fuzzy notion of "world as it should and could be" (Mitchell et al. 2020, p. 4), we need to be able to articulate what it is we find undesirable and what we want to produce in its stead. Rather than treating bias as a problem that can removed from the world, we need to more directly consider what kind of world we want to create, and to this end we are better off relying on conceptualizations of inequality, structured distributions, and power.

Over the previous decade, a large body of work in critical algorithm studies (Gillespie and Seaver 2016) and critical data studies (Iliadis and Russo 2016) has emerged to address the emerging structures of our digitally-mediated interactions and their relations of power, alongside a rapid development of fairness and ethics as key concerns in AI and ML research (Jobin et al. 2019; Mitchell et al. 2020). Black AI scholars, including Timnit Gebru and Ruha Benjamin, have helped move such critiques from the field's margins and into greater visibility, and various radical approaches are actively being developed in opposition to AI's conservative tendencies (Cifor et al. 2019; Costanza-Chock 2020; Kalluri 2020). Sociological contributions to these discussions have largely filtered through the field of science and technology studies (STS), or tackled questions of governance, discourse, AI agency, and social interaction (for example, Roberge and Castelle 2021). Ruha Benjamin's (2019) work stands out for its analysis of racial inequality and social structure, but as demonstrated in a recent controversy involving an ML system that 'upsampled' non-white faces (such as Barack Obama's) into white faces (Johnson 2020b), many AI researchers and developers have remained unfamiliar with this work—content to see 'the data' as a source of bias, or treating underlying social inequalities as out of scope for the field. Therefore, this article has begun by examining what bias means for AI practitioners, and will now proceed to reconceptualize the perpetuation of bias in more interdisciplinary and sociological terms.

To look beyond 'the data', interdisciplinary approaches to ethical issues in AI can situate these problems in relation to larger historical and social forces, but the challenge is to turn these into a positive way forward. To this end, social science scholarship gives us language with which we can better specify issues and how to address them. To have meaningful impact, this interdisciplinary engagement needs to happen early in the development of AI systems; it is not simply a matter of adding the missing social context to an already-formulated problem. When we begin by naming and analyzing the social structures we find problematic, we can think about ways of changing them or addressing their harms. The later in the development process that this happens, the more likely that the socio-technical system will already be part of the problem, and reforms can only be superficial. The full benefit of interdisciplinarity can only be realized by starting with the fundamental question of what a desirable outcome looks like, given what we know about how inequality is organized. From there, we can work towards concrete steps for action, but this must be based on a thorough understanding of a particular domain of inequality, and with the involvement of those who are imagined to benefit from the outcome.

3.1 Desirable outcomes: conceptualizing bias as inequality

To reiterate the previous, there are two basic definition of bias in AI, as being (1) inaccurate or (2) undesirable. These two forms of bias may be related, but they should not be conflated; inaccuracy is generally undesirable, but it is possible for a prediction to be accurate and also have harmful, undesirable effects. The following relates to the second definition

of bias as an unwanted obstacle to a better world, rather than the accuracy of the algorithm. Within AI scholarship, the fuzzy goal of doing "social good" has predominated, but by not articulating a specific notion of what is good, AI systems often end up further entrenching social harms (Green 2019). This is where interdisciplinary insights have the most to offer, as an alternative to the removal of bias and a way towards a more positive vision.

The most obvious way that we can move beyond the negative orientation of 'removing bias' is to specify social inequality as the problem, and equality or equity as a desirable outcome to work towards. However, it is far from self-evident what these terms mean in a normative sense. Is it a society with a more equitable distribution of resources and wealth? Is it a society, where individuals have greater autonomy or equality in their opportunities to obtain unequal rewards? To what extent should our normative horizon be focused on individual equality, as opposed to how groups and collectives relate to one another?

There are different ways of designing algorithms to maximize equality given these problems, but any assumptions made in operationalizing fairness have political implications. This is because "algorithmic fairness… reproduces, at a technical level, the tensions inherent in the political philosophy of justice" (Hacker 2018, p. 1183), such as the tension between individual and group-level fairness, anticlassification and antisubordination theory (Barocas and Selbst 2016), or the related tension between equality-of-opportunity and substantive equality (Baumann and Rumberger 2018). While equality does depend upon some sort of mutual recognition of commonality, the nature of this shared character (the way in which we are supposedly equal), and the idea of a social order based on equality has been invented and developed along different lines (Rosanvallon 2013). The goal of equality does not exist in some abstract political and theoretical space, but in relation to specific power structures and power struggles.

For example, the abstract term 'social inequality' is most often used to refer to economic inequality, either between social classes, or some conceptual alternative relating to wealth and socio-economic status (Grusky 2014). The literature on bias in AI has relatively little to say about economic inequality, other than as an outcome of algorithmic decision-making. 'Class bias' is a term that is more likely to appear in sociology (see Goetz 1997) than AI ethics, where the focus is typically on discrimination against protected categories such as race, gender, and disability (Costanza-Chock 2020). Bias can certainly be a concern when inaccuracies and errors have economic impacts, such as when some people are wrongly excluded or denied credit because of errors in their credit reports. However, when people are included for targeted discrimination or exploitation, these processes tend to be discussed as "predatory" rather than biased (see Johnson et al. 2019). Conceptualizing these processes as predation or bias calls out the most exploitative dimension of capitalism, but this is not a foundation on which to proceed to a better society. A bias-free world could be one where every person is equally surveilled and controlled, or equally targeted by predatory lenders.

While we could characterize algorithms that target the vulnerable and economically precarious as being biased against the poor, it is more accurate to say that the poor are managed and controlled, based on longstanding assumptions that it is desirable to discriminate against the poor (Eubanks 2017), or that the poor are exploited by algorithms designed to "target and fleece the population most in need" (O'Neil 2016, p. 81). We therefore need to examine the structures through which the poor are kept poor, so that we can eliminate or circumvent these where possible, and strengthen the sorts of structures that could promote social mobility and economic betterment. Class, like other dimensions of inequality, is reproduced on an ongoing basis (Grusky 2014), preserving the privileges of wealth as well as the disadvantages of poverty from one generation to the next. By the time AI systems enter the picture, this unequal ground truth is well-established, intersecting with other dimensions of inequality like gender, race, and disability (Hoffmann 2019).

Building a fairer decision-making algorithm will only have a superficial impact on fundamental inequalities, but a structural analysis can identify, where more systemic changes would be effective, including where automated decision-making should be removed from a process (Eubanks 2017). A radical approach goes further in reducing unequal outcomes and unequal distributions of resources, goods, or other kinds of "holdings" (Segall 2013). This also broadens the focus from disadvantages or obstacles, to considering the advantages or privileges maintained by existing structures. Instead of imagining that we can lift the marginalized up to become full participants in some 'level playing field', radical arguments often seek to overturn the competitive playing field altogether to pursue other forms of social relations.

These remain very general considerations for reconceptualizing bias as inequality, and a great deal will depend on the kinds of inequalities with which we are concerned and how these intersect in a given context. Therefore, I will next turn to the more specific issue of racial inequality in the context of settler colonialism. Just as we cannot rely on a generalized approach to all forms of inequality, racism and racial inequality are not universal phenomena that can be separated from their socio-historical circumstances (see Lentin 2020). 'Racial bias' does not exist in the abstract, general sense, but only within specific social formations – namely, the different social structures built through colonial processes.

4 Race and colonialism

Racial bias is one of the most prominent forms of bias discussed in AI and ML, with the example of the COMPAS recidivism prediction algorithm having achieved a paradigmatic status in discussions of the topic (Wong 2020). Rather than revisiting that well-worn example, I want to broaden the discussion by introducing a related case of algorithmic racial bias from a non-U.S. context. In Canada, criminal risk assessments methods have also been critiqued for subjective and systemic biases against Black and Indigenous prisoners, whose high level of incarceration is considered a problem of 'overrepresentation' in comparison to the larger population. Canadian recidivism prediction has not been automated, but remains based on "pen and paper" bureaucratic procedures, actuarial scales, and human judgment (Cardoso 2020). Much as COMPAS was developed as a way to replace human judgment with assessments based on objective statistics, we can ask whether better statistical models and ML algorithms could remove racial bias from the equation in Canada, given the well-documented issues in Canadian criminal justice.

Certainly, we can imagine how an AI system could address the problem by automating risk assessment, but as with reforming COMPAS, the consequences are likely to be conservative. We can either replace the algorithm or alter its biases, both of which might be improvements over a flawed algorithm, but these remain "reformist reforms" (Green 2019) to the criminal justice system, leaving the inequalities it perpetuates largely intact. In other words, if our interventions happen where an AI system is introduced to make an existing decision-making process more efficient, accurate, or fair, the scope of any changes we can hope to effect will be extremely limited. The alternative approach (namely, abolition) has been articulated by critical AI scholars (most notably, Benjamin 2019) whose work can lead us to an entirely different and earlier starting place. If our focus is racial inequality, that means starting from an understanding of how race exists as a social structure, using this understanding to consider racial inequality in a specific context, and then seeing how this structure can most effectively be changed. Only then should we consider the role AI systems could have (if any), and only as long as we are working with the people who are ultimately supposed to benefit from social change.

If racial inequality is the problem we want to address, the first step is abandoning the notion that this inequality is the result of racism in the form of an irrational, individual human bias that can be objectively automated out of existence. This approach is the result of a popular contemporary view of race, which imagines race to be a neutral category and treats racism as an individual pathology (Lentin 2020). But racism does not simply manifest as individual people making biased decisions—it involves the configuration and reproduction of social structures that create inequalities. To have a sense of what this means, we have to begin by consulting the relevant literature.

There are significant theoretical affinities as well as differences among scholars of Black radical thought, critical race theorists, neo-Marxists, and sociologists of race, regarding concepts such as white supremacy, racialization, and the relevance of class (Robinson 2000; Omi and Winant 2014; Bonilla-Silva 2015; Walton 2020). Despite their disagreements, these approaches share a view of race and racism that is grounded in social structure—explanations that go beyond individual human biases and racist intent to the systematic ways that racial hierarchies are maintained. This theoretical background is crucial to understanding how AI technologies are implicated in the 'tech-to-prison pipeline', and how certain forms of AI can only strengthen these structures and their oppressive effects (Coalition for Critical Technology 2020; Hanna et al. 2019).

Racial categories are the outcomes of historically and culturally-contingent classification schemes that associate hierarchical values with supposedly natural groups of people (Benthall and Haynes 2019). If there is a common thread to contemporary understandings of race around the world, it is their origin in an "imperial imperative" to classify subject populations (Hacking 2005), which, as a consequence of five hundred years of European imperialism, has led to global inequalities between white and non-white people (Mills 1997). Far from being an immaterial fiction, race is used to materially inscribe inequalities onto human bodies and societies, operating as a "technology for the management of human difference, the main goal of which is the production, reproduction, and maintenance of white supremacy on both a local and a planetary scale" (Lentin 2020, p. 11). Because it is so deeply rooted, this inequality manifests as 'bias' in so many ways, whether as non-white individuals targeted for increased social control (Browne 2015) or upsampled into white faces (Johnson 2020b). This is also why it is more effective to attack racial inequality closer to the root, rather than a late-stage decision-making algorithm such as a recidivism predictor.

Because race and colonialism are so closely linked, theories of colonialism provide a valuable means of understanding issues in AI through historical inequalities and asymmetrical power relations (Mohamed et al. 2020), and which are also helpful in moving beyond approaches to AI ethics grounded in formalizing fairness and equality. The fight for equality has been historically important in anti-racist and anti-colonial political struggles, but so has the radical goal of dismantling systems of oppression and the creation of new sources of solidarity. The sorts of inequalities that exist in settler colonial societies (including those

in the Americas, as well as Australia and New Zealand, see Walter et al 2020) cannot be addressed by reducing bias in decision-making. In a society such as Canada for instance, the 'unequal ground truth' is one in which Indigenous populations may be governed by different institutions, receiving inferior educational opportunities, inferior access to health care and basic infrastructure, and where Indigenous children are much more likely to be raised in poverty and foster care, as consequences of generations of colonization (Truth and Reconciliation Commission of Canada 2015; Blackstock 2017; Statistics Canada 2019). The common cause of these inequalities is not individual, societal or historical bias, but settler colonialism. This is now a well-theorized social structure, organized around dispossession, Indigenous elimination, and control over land (Veracini 2010). Putting a name to the structures of inequality moves us closer to understanding the otherwise amorphous social context, why unfairness and injustice are distributed as they are, and towards normative approaches that go beyond equality.

4.1 Racial inequality and positive goals

If we see our challenge as one of minimizing bias, we might try to reduce group differences, where inequalities exist, working towards similar treatment and social integration (for example, Benthall and Haynes 2019). However, social integration and universal treatment cannot be assumed as goals when addressing racial and ethnic inequality, especially in a colonial context, where identical treatment can amount to assimilation. In Canada, to assume that the ideal is universal citizenship with equal rights and opportunities for all would ignore those Indigenous peoples who may not identify as Canadians, and who possess different rights through treaty obligations. Differential treatment is institutionalized in Canada's justice system, where a one-size-fits-all approach to risk assessment is known to produce unequal harms (Cardoso 2020), and where distinct legal considerations are applied to Indigenous persons to address the effects of colonialism. Rather than equality, Canada's officially-recognized objective is 'reconciliation' (Truth and Reconciliation Commission of Canada 2015), and more radical scholarship asks us to consider ways of working towards decolonization and Indigenous resurgence (Simpson 2016).

Once again, by naming desired goals, or articulating a positive value to work towards, we can be more specific than the negative process of removing bias. In doing so, it is important that our normative goals are attuned to social context and the goals of the people being affected, rather than universal notions of equality. Both decolonization and reconciliation involve recognizing the distinctiveness of Indigenous peoples, and the legacy of injustices imposed on them and legitimated by the state. Key among these are the abrogation of Indigenous sovereignty, the dispossession of land, and the purposeful destruction of Indigenous cultures. As with other forms of social inequality, settler colonialism is actively reproduced through everyday practices, distinctions, and ideas (Veracini 2010). There is no technological 'fix' for colonialism, dispossession of land, and loss of culture, but we can still consider how AI systems "reproduce colonial ontology and epistemology" (Costanza-Chock 2020, p. 67), and what these systems would look like if they were designed to support Indigenous sovereignty and resurgence.

4.2 Designing technologies in an unequal world

For an AI developer working on any project, where there is a concern about bias, the concrete steps that can be taken will depend on the extent of one's ability to shape the project and its objectives. If the goal is to achieve better performance on some classification or prediction task, the system's political consequences will likely be quite conservative, preserving existing distinctions and hierarchies. But regardless of the problem that AI is being used to address, wherever racial bias (for example) is an issue, the least that a developer can do is to understand what race is, and how racial inequality is structured in society. While this might seem like an obvious point, there is still an enormous amount of work being done in computing and data science to classify races, genders, emotional states, or potential for criminality, with only the shallowest ontological engagement with these phenomena and what is known about them in other fields (Barrett et al. 2019; Hanna et al. 2019; Scheuerman et al. 2019). My argument is that AI researchers and developers need to be able supplant a term like 'racial bias', which restricts further analysis, with theories of racial inequality that open up further avenues for analysis—including examining how race intersects with other social hierarchies (Hoffman 2019). Doing so makes it possible to specify goals or values other than accuracy, efficiency, equality, fairness, or reducing bias. It enables us to evaluate the extent to which an AI system actually makes the world a better place, as judged in comparison to the kind of world we want to create while considering the extent to which these systems continue to reproduce the existing social order.

A more open-ended opportunity comes from being able to tackle forms of social inequality as a problem, with the freedom to work towards whatever best achieves positive change. This means that 'doing good' is not an afterthought to some existing process, but the starting point, which again involves specifying exactly what kind of 'good' we are trying to achieve. The overrepresentation of racialized groups in prisons is not going to be fixed by an algorithm that is race-blind, or designed to imprison people at representative rates; it is going to require the sorts of political changes that get us closer to the ideals of abolition, decolonization, or resurgence. For a technologist, this requires understanding

that many well-intentioned efforts to do good through technology have produced harm (Green and Viljoen 2020), and that there are many problems that we cannot design our way out of. As Benjamin (2019) argues, design thinking "colonizes" other forms of human activity, and she suggests that "maybe what we must demand is not liberatory *designs* but just plain old liberation" (p. 179). But Benjamin does not dismiss the potential that new technologies have in shaping the world or our experience of it, just that we should not privilege design as a form of political action.

If technology can indeed be used for liberation or reordering power relations, radical approaches to technology begin at the bottom or the margins, are attuned to the needs of communities that have been historically disempowered or excluded from decision-making, and recognize that new technologies are often not required to meet the most important needs. To this end, "design justice", as articulated by Costanza-Chock "requires full inclusion of, accountability to, and ultimately control by people with direct lived experience of the conditions the design team is trying to change" (2020, p. 99). For issues impacting Indigenous populations, this means centering Indigenous concerns and perspectives (see Lewis et al 2020), including the right to exercise sovereign control over data produced by or about Indigenous peoples (Walter et al 2020). While self-determination and sovereignty have specific relevance for Indigenous peoples, the more broadly applicable radical values for design prioritize agency and involvement for those directly impacted by technologies, producing "AI that is faithful to the needs of data subjects and allows them to opt out freely" (Kalluri 2020). As AI is increasingly used to make decisions for and about people, it is these people and their needs that should steer the path of AI development.

5 Conclusion

The larger lesson here is the importance of attending to the specifics of inequality and social structure. In short, fairness algorithms are not generalizable and social inequalities are deep and multiple. The concept of bias is limiting and should often jettisoned, where more specific conceptualizations of inequality are available. Rather than being concerned over how socio-technical systems reproduce pre-existing biases, we can actually name what we want to avoid reproducing: identifying processes, structures, hierarchies and concepts that have already been articulated by critical AI scholars and those in the social sciences and humanities. Being specific also helps us to name desirable alternatives to reproducing injustice, and orients us to where our actions can have meaningful impact.

Conservative approaches to AI ethics seek to achieve equal treatment under existing institutions, but enabling radical change (such as alternatives to prison and policing) is a tougher problem. A first step is appreciating that we are dealing with political rather than technical problems, which cannot be solved by better models and AI systems (Green 2018; Wong 2020). Given that politics is fundamentally about power, we would do well to recognize how these systems currently work to intensify, maintain, and optimize existing forms of power. Critical scholarship provides the tools to understand how unjust and oppressive structures are upheld; racism, sexism, and ableism exist as processes rather than individual traits, and are maintained by the daily 'unintentional' actions of well-meaning people. Given our social circumstances, one cannot simply opt out of patriarchy, colonialism, or capitalism, but understanding how these structures work and are sustained can inform ways to limit our support, including through refusal (Cifor et al. 2019) and the work we choose not to do. The more difficult task, of actually shifting power (Kalluri 2020) and producing alternatives to existing injustices, may involve technical innovations based on different 'use cases' or ontologies, but these are often among the least effective ways of pursuing political change.

It is very difficult to make the world a better place with an algorithm, and almost inevitable that automated decisions will promote distinctions that cause harm, simply through the reproduction of harmful, preexisting inequalities. This is the conservative tendency of AI systems, which has been called the reproduction of societal bias. I have argued that this bias is better understood as the intersection of different structures of inequality, as named and analyzed by scholars in the social sciences and humanities prior the current era of machine learning. These theories have informed some of the foundational contributions of critical AI scholars, which have unfortunately remained marginal to the field of AI development. A great deal of work in computing and data science continues to discriminate between social categories, without seriously engaging with what is known about these categories and their relationships in other disciplines. Such interdisciplinary engagement can sometimes inform the design of AI systems to make them less harmful, but the goal of making the world a better or fairer place requires a great deal more. To this end, being able to name the structures we find problematic has the benefit of providing positive goals to work towards, rather than just negating bias. However, existing applications of AI often have limited potential to make progress towards these goals, which are inherently political, and are often better served through more conventional kinds of political action. The problem of overrepresentation in the criminal justice system is not going to be solved by a fairer algorithm, but by steps to address deeper injustices and processes of criminalization. An AI system is more likely to reproduce underlying inequalities than to radically transform them, but this conservative tendency can

be understood and explained using language that is specific to a given context of inequality, providing a step towards formulating radical alternatives.

Acknowledgements This is to acknowledge that an earlier draft of this article is available on arXiv: 200708666 [cs] https://arxiv.org/abs/2007.08666.

Funding Research for this article was supported by funding from the University of British Columbia.

Compliance with ethical standards

Conflict of interest The authors declared that they have no conflict of interest.

References

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104:671–732

Barrett LF, Adolphs R, Marsella S et al (2019) Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. Psychol Sci Public Interest 20:1–68. https://doi.org/10.1177/1529100619832930

Baumann E, Rumberger JL (2018) State of the Art in Fair ML: From Moral Philosophy and Legislation to Fair Classifiers. ArXiv181109539 Cs Stat http://arxiv.org/abs/1811.09539

Benjamin R (2019) Race after technology: abolitionist tools for the new Jim code. Polity Press, Cambridge, U.K.

Benthall S, Haynes BD (2019) Racial categories in machine learning. Proc Conf Fairness Account Transpar FAT 19:289–298. https://doi.org/10.1145/3287560.3287575

Birhane A (2020) Fair Warning. In: Real Life. https://reallifemag.com/fair-warning/. Accessed 29 December 2020

Blackstock C (2017) Reflections on reconciliation after 150 years since confederation: an interview with Dr Cindy Blackstock. Ott Law Rev 49:13–28

Bonilla-Silva E (2015) The structure of racism in color-blind, "post-racial" America. Am Behav Sci 59:1358–1376. https://doi.org/10.1177/0002764215586826

Bourke R (2018) What is conservatism? History, ideology and party. Eur J Polit Theory 17:449–475. https://doi.org/10.1177/1474885118782384

Browne S (2015) Dark matters: on the surveillance of blackness. Duke University Press, Durham

Cardoso T (2020) Bias Behind Bars: A Globe investigation finds a prison system stacked against Black and Indigenous inmates. In: Globe and Mail. https://www.theglobeandmail.com/canada/article-investigation-racial-bias-in-canadian-prison-risk-assessments/. Accessed 26 Oct 2020

Carr N (2014) The Limits of Social Engineering. In: MIT Technol. Rev. https://www.technologyreview.com/2014/04/16/173156/the-limits-of-social-engineering/. Accessed 29 Jun 2020

Chiang T (2017) Silicon Valley Is Turning Into Its Own Worst Fear. In: BuzzFeed News. https://www.buzzfeednews.com/article/tedchiang/the-real-danger-to-civilization-isnt-ai-its-runaway. Accessed 14 Jul 2020

Cifor M, Garcia P, Cowan TL, et al (2019) Feminist data manifest-no. https://www.manifestno.com/home. Accessed 31 Jan 2021

Coalition for Critical Technology (2020) Abolish the #TechToPrisonPipeline. In: Medium. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16. Accessed 3 Jul 2020

Cooley M (1995) The myth of the moral neutrality of technology. AI Soc 9:10–17. https://doi.org/10.1007/BF01174475

Corbett-Davies S, Goel S (2018). The measure and mismeasure of fairness: a critical review of fair machine learning. ArXiv1808.00023 Cs http://arxiv.org/abs/1808.00023

Costanza-Chock S (2020) Design justice: community-led practices to build the worlds we need. MIT Press, Cambridge, MA

Eubanks V (2017) Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York

Friedler SA, Scheidegger C, Venkatasubramanian S (2016) On the (im)possibility of fairness. ArXiv160907236 Cs Stat http://arxiv.org/abs/1609.07236

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Trans Inf Syst 14:330–347. https://doi.org/10.1145/230538.230561

G7 Science Academies (2019) Artificial intelligence and society. https://rsc-src.ca/sites/default/files/Artificial%20intelligence%20and%20society%20G7%202019.pdf. Accessed 31 May 2019

Gillespie T, Seaver N (2016) Critical Algorithm Studies: a Reading List. In: Soc. Media Collect. http://socialmediacollective.org/reading-lists/critical-algorithm-studies/. Accessed 7 Jul 2020

Goetz B (1997) Organization as class bias in local law enforcement: arson-for-profit as a "nonissue." Law Soc Rev 31:557–588

Green B (2018) Data science as political action: grounding data science in a politics of justice. ArXiv181103435 Cs https://arxiv.org/abs/1811.03435

Green B (2019) "Good" isn't good enough. In: proceedings of the AI for Social Good workshop at NeurIPS. Vancouver. https://www.benzevgreen.com/wp-content/uploads/2019/11/19-ai4sg.pdf. Accessed 29 December 2020

Green B, Hu L (2018) The myth in the methodology: towards a recontextualization of fairness in machine learning. https://scholar.harvard.edu/files/bgreen/files/18-icmldebates.pdf. Accessed 1 October 2020

Green B, Viljoen S (2020) Algorithmic realism: expanding the boundaries of algorithmic thought. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, Barcelona, Spain 19–31

Grusky DB (2014) Social stratification: class, race, and gender in sociological perspective, 4th edn. Routledge, New York

Hacker P (2018) Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU law. Common Mark Law Rev 55:1143–1185

Hacking I (2005) Why Race Still Matters. Daedalus 134:102–116

Hanna A, Denton E, Smart A, Smith-Loud J (2019) Towards a critical race methodology in algorithmic fairness. ArXiv: 1912.03593 Cs http://arxiv.org/abs/1912.03593

Hartley S (2017) The Fuzzy and the Techie: Why the Liberal Arts Will Rule the Digital World. Houghton Mifflin Harcourt, Boston

Hoffman SG (2017) Managing ambiguities at the edge of knowledge: research strategy and artificial intelligence labs in an era of academic capitalism. Sci Technol Hum Values 42:703–740. https://doi.org/10.1177/0162243916687038

Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inf Commun Soc 22:900–915. https://doi.org/10.1080/1369118X.2019.1573912

Iliadis A, Russo F (2016) Critical data studies: an introduction. Big Data Soc 3:1–7. https://doi.org/10.1177/2053951716674238

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson K (2020a) NeurIPS requires AI researchers to account for societal impact and financial conflicts of interest. VentureBeat.

https://venturebeat.com/2020/02/24/neurips-requires-ai-researchers-to-account-for-societal-impact-and-financial-conflicts-of-interest/. Accessed 29 Dec 2020
Johnson K (2020b) AI Weekly: A deep learning pioneer's teachable moment on AI bias. VentureBeat. https://venturebeat.com/2020/06/26/ai-weekly-a-deep-learning-pioneers-teachable-moment-on-ai-bias/. Accessed 27 Jun 2020
Johnson K, Pasquale F, Chapman J (2019) Artificial intelligence, machine learning, and bias in finance: toward responsible innovation. Fordham Rev 88:499–529
Kalluri P (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature 583:169–169. https://doi.org/10.1038/d41586-020-02003-2
Kusner MJ, Loftus JR (2020) The long road to fairer algorithms. Nature 578:34–36. https://doi.org/10.1038/d41586-020-00274-3
Leith P (1990) Formalism in AI and computer science. Ellis Horwood, New York
Lentin A (2020) Why race still matters. Polity Press, Cambridge
Lepri B, Oliver N, Letouzé E et al (2018) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 31:611–627. https://doi.org/10.1007/s13347-017-0279-x
Lewis JE, Abdilla A, Arista N, et al (2020) Indigenous Protocol and Artificial Intelligence Position Paper. https://spectrum.library.concordia.ca/986506/. Accessed 15 Oct 2020
Lukes S (2005) Power: a radical view, 2nd edn. Red Globe Press, London
Lury C (2018) Introduction: Activating the present of interdisciplinary methods. In: Lury C, Fensham R, Heller-Nicholas A et al (eds) Routledge handbook of interdisciplinary research methods. Routledge, New York, pp 1–25
Machin A, Stehr N (2016) Inequality in Modern Societies: Causes, Consequences and Challenges. In: Stehr N (ed) Machin A. Understanding Inequality, Social Costs and Benefits. Springer, pp 3–34
Mills CW (1997) The racial contract. Cornell University Press, Ithaca, NY
Mitchell S, Potash E, Barocas S, et al (2020) Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. ArXiv181107867 Stat http://arxiv.org/abs/1811.07867
Mohamed S, Png M-T, Isaac W (2020) Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos Technol 33:659–684. https://doi.org/10.1007/s13347-020-00405-8
Nath R, Sahu V (2020) The problem of machine ethics in artificial intelligence. AI Soc 35:103–111. https://doi.org/10.1007/s00146-017-0768-6
Nisbet RA (1952) Conservatism and sociology. Am J Sociol 58:167–175
O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown, New York
Omi M, Winant H (2014) Racial Formation in the United States, 3rd edn. Routledge, New York
Penn J (2018) AI thinks like a corporation—and that's worrying. The Economist. https://www.economist.com/open-future/2018/11/26/ai-thinks-like-a-corporation-and-thats-worrying. Accessed 29 Dec 2020
Roberge J, Castelle M (eds) (2021) The cultural life of machine learning: an incursion into critical AI studies. Springer International Publishing, Cham
Robinson CJ (2000) Black marxism: the making of the black radical tradition, 2nd edn. University of North Carolina Press, Chapel Hill
Rosanvallon P (2013) The society of equals. Harvard University Press, Cambridge, Massachusetts
Scheuerman MK, Paul JM, Brubaker JR (2019) How Computers see gender: an evaluation of gender classification in commercial facial analysis services. Proc ACM Hum-Comput Interact 3:1–33. https://doi.org/10.1145/3359246
Segall S (2013) Equality and opportunity. Oxford University Press, Oxford
Selbst AD, boyd d, Friedler SA, et al (2019) Fairness and abstraction in sociotechnical systems. In: proceedings of the conference on fairness, accountability, and transparency. ACM, New York 59–68
Shah D, Schwartz HA, Hovy D (2019a) Predictive biases in natural language processing models: a conceptual framework and overview. ArXiv191211078 Cs http://arxiv.org/abs/1912.11078
Shah R, Gundotra N, Abbeel P, Dragan AD (2019b) On the feasibility of learning, rather than assuming, human biases for reward inference. ArXiv190609624 Cs https://arxiv.org/abs/1906.09624
Silberg J, Manyika J (2019) Notes from the AI frontier: tackling bias in AI (and in humans). https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.ashx. Accessed 29 Dec 2020
Simpson LB (2016) Indigenous resurgence and co-resistance. Crit Ethn Stud 2:19–34. https://doi.org/10.5749/jcritethnstud.2.2.0019
Statistics Canada (2019) Statistics on Indigenous peoples. https://www.statcan.gc.ca/eng/subjects-start/indigenous_peoples. Accessed 21 Aug 2019
Suresh H, Guttag JV (2020) A Framework for Understanding Unintended Consequences of Machine Learning. ArXiv190110002 Cs http://arxiv.org/abs/1901.10002
Truth and Reconciliation Commission of Canada (2015) Honouring the truth, reconciling for the future: summary of the final report of the truth and reconciliation commission of Canada
Veracini L (2010) Settler colonialism: a theoretical overview. Palgrave Macmillan, Houndmills, Basingstoke
Walter M, Kukutai T, Carroll SR, Rodriguez-Lonebear D (eds) (2020) Indigenous data sovereignty and policy. Routledge, London
Walton S (2020) Why the critical race theory concept of 'White supremacy' should not be dismissed by neo-Marxists: lessons from contemporary Black radicalism. Power Educ 12:78–94
West SM, Whittaker M, Crawford K (2019) Discriminating systems: gender, race and power in AI. AI now institute. https://ainowinstitute.org/discriminatingsystems.pdf. Accessed 31 Jan 2021
Whittaker M, Crawford K, Dobbe R, et al (2018) AI Now 2018 report. AI now institute. https://ainowinstitute.org/AI_Now_2018_Report.pdf. Accessed 31 Jan 2021
Winner L (1980) Do artifacts have politics? Daedalus 109:121–136
Wolfe AB (1923) Conservatism and radicalism: some definitions and distinctions. Sci Mon 17:229–237
Wong P-H (2020) Democratizing algorithmic fairness. Philos Technol 33:225–244. https://doi.org/10.1007/s13347-019-00355-w