Algorithmic Harms and Algorithmic Wrongs

Nathalie Diberardino
Western University
ndibera@uwo.ca

Clair Baleshta
Western University
cbalesht@uwo.ca

Luke Stark
Western University
cstark23@uwo.ca

ABSTRACT

New artificial intelligence (AI) systems grounded in machine learning are being integrated into our lives at a rapid rate, but not without consequence: scholars across domains have increasingly pointed out issues related to privacy, transparency, bias, discrimination, exploitation, and exclusion associated with algorithmic systems in both public and private sector contexts. Concerns surrounding the adverse impacts of these technologies have spurred discussion on the topics of algorithmic harm. However, the overwhelming majority of articles on said harms offer no definition as to what constitutes 'harm' in these contexts. This paper aims to address this omission by introducing one criterion for a suitable account of algorithmic harm. More specifically, we follow Joel Feinberg in understanding *harms* as distinct from *wrongs*, where only the latter necessarily carry a normative dimension. This distinction highlights issues in the current scholarship surrounding the conflation of algorithmic harms and wrongs. In response to these issues, we put forth two requirements for upholding the harms/wrongs distinction when analyzing the increasingly far-reaching impacts of these technologies and suggest how this distinction can be useful in design, engineering, and policymaking.

CCS CONCEPTS

• **Social and professional topics** → Computing / technology policy.

KEYWORDS

algorithmic harms, algorithmic wrongs, artificial intelligence, bias, fairness, artificial intelligence, design, policy

ACM Reference Format:
Nathalie Diberardino, Clair Baleshta, and Luke Stark. 2024. Algorithmic Harms and Algorithmic Wrongs. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3630106.3659001

1 INTRODUCTION

Recent scholarship has tied concerns surrounding such AI systems' increasingly disruptive societal impacts to the concept of *harm* [48,52]. Harm—how it should be defined, understood, and prevented—has been a topic of intense academic interest for some time [22,60,63]. In the context of novel digital technologies, scholars have identified various features and categories of algorithmic harm [13,48,71]. Concepts closely related to harm such as violence [30,31,32,58] and injustice [8] are also frequently mobilized to interrogate AI systems and digital technologies more broadly. However, despite a significant and growing discourse surrounding the harms of such technologies, there has been little explicit analysis of the concept of "harm" itself. The overwhelming majority of articles on algorithmic harm offer no definition as to what constitutes it. Those scholars who do define harm [49,76,77] typically rely on a comparative account by legal philosopher Joel Feinberg, wherein harms are understood as setbacks to interests [22]. This lack of nuance in the analysis of 'harm' itself raises a number of questions, as our intuitive notions of harm are often prone to biases, inconsistencies, and other inaccuracies [9,11]. Perhaps most crucially, existing scholarship has overlooked an important distinction regarding harm made by Feinberg in his work: for Feinberg, a harm is a non-normative concept that is categorically different from a moral wrong.

In this paper, we argue that the conceptual distinction between harms and wrongs is analytically important for crafting technical and policy responses to the deleterious effects of AI systems. Accounts of algorithmic harms in the context of artificial intelligence technologies seeking to highlight the unjust effects of such systems would benefit strongly from acknowledging and making use of this conceptual distinction. In Section 2, we begin by providing a brief review of the current literature on algorithmic harms, including how harm is most often conceptualized and understood in digital contexts. In Section 3, we note that despite this widespread attention to harm in relation to algorithmic technologies, most of the work in this literature does not explicitly define the concept of harm. In doing so, we engage with Feinberg's popular philosophical theory of harm, emphasizing his underrecognized yet key distinction between the conceptual nature of harms and wrongs. In Section 4, we argue that recognizing this distinction is especially crucial in the context of artificial intelligence technologies to grasp the scope and nature of algorithmic harms, especially given their increasing ubiquity. Finally, in Section 5 we articulate the utility of distinguishing between harms and wrongs, both in design and engineering and in policymaking. We also put forth two requirements that need to be met within both scholarship and design practice in order to help uphold the harms/wrongs distinction. First, we argue that respecting the distinction requires us to understand algorithmic harms and algorithmic wrongs as contextual and socially embedded in nature. Second, we suggest that engaging with a diversity in community and stakeholder perspectives is necessary to track Feinberg's distinction and ultimately appropriately conceptualize, and respond, to the impacts of new AI technologies.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659001

2 PERSPECTIVES ON ALGORITHMIC HARMS

Despite optimism regarding the alleged utility and positive social impacts of new algorithmic technologies, the increasing ubiquity of these systems has evoked greater attention to how they might be *harmful* in the various contexts they are being used. For example, the uptake of artificial intelligence technologies in specific domains like healthcare, education, and policing has spurred numerous analyses of their potential harm. In healthcare, stakeholders have raised concerns about the potential for AI-based diagnostic technologies to enable inaccurate judgements or exacerbate existing racial inequalities in clinical care provision [1,56]. In policing, the use of predictive tools such as PredPol have raised alarm regarding these technologies role in exacerbating racist carceral policies [57]. There is also significant attention surrounding the harmful impacts of AI systems on the spread of misinformation and consequent threats to democratic institutions [85]. In addition, the potential harms associated with the use of algorithmic systems in contexts like education, university admissions, hiring, insurance, welfare, and credit rating have all been widely recognized [18,19,21,47,86]. Other prominent topics in the critical AI literature, such as privacy, responsibility, and bias are often closely related to concerns surrounding harm [15,84,86]. Issues surrounding bias, for instance, are often explained in terms of the harms resulting algorithmic discrimination, exclusion, and exploitation [46,51,62]. As such, the idea of *algorithmic harm* has become a central focus in the interdisciplinary literature on AI ethics in response to this growing body of evidence for AI systems' malign impacts [26,64].

Many scholars have begun to qualify harm, highlighting various characteristics of it as it appears in algorithmic contexts. Some have distinguished between different kinds of algorithmic harm. Anna Lauren Hoffmann [29] for instance, charts the distinction between 'distributional' and 'dignitary harms,' where the former pertains to injustices in the way society distributes "rights, responsibilities, and resources," while the latter concerns affronts to an individual's self-respect. Hoffmann emphasizes that "self-respect confers upon individuals a sense of their own value and a conviction that their plans of life are worth pursuing" [29:81–82]. A dignitary harm occurs when a system undermines this self-respect of the user or otherwise produces unequal social or relation effects on society [9,23]. Scholars have also discussed terms closely related to harm in their efforts to understand the impacts of digital technologies. Hoffmann [32], for instance, also introduces the idea of data violence, defined as "the material, symbolic, and other violences inflicted by and through data technologies and their purveyors" [32:2]. One such form is "discursive violence," which, she says, "operates by diffusing resistance, deepening dependency on oppressive structural conditions, and preserving the potential for other forms of violence, including physical, material, and symbolic violences." Discursive violence diffuses resistance by normalizing conditions that "make other (material, symbolic) violences look right, or at least not wrong" [32:5]. Mimi Onuoha defines "algorithmic violence" as the way in which an algorithm or automated decision-making system inflicts [violence] by preventing people from meeting their basic needs" [58]. Harm is also sometimes conceptualized as an injustice, whereby marginalized and oppressed groups are often disproportionately impacted by new data practices and algorithmic systems [7,8,16,51]. Other ways of conceptualizing particular kinds of algorithmic harms include terms like "the automation of virtue" and "moral deskilling" to highlight the danger that algorithms will deny us the experiences of making and ability to make our own moral choices and develop character and moral virtue [40,83]. Likewise, the notion of "algorithmic pollution" is a way to denote the "unjustified, unfair, discriminatory, and other harmful effects of automated algorithmic decision-making" for individuals, groups, and societies [44:9].

Yet despite this growing literature on algorithmic harm, there has been little elucidation as to what exactly constitutes a "harm" as applied to discussions of digital technologies, evidenced by the fact that most articles that discuss AI or algorithmic harm do not provide any definition of harm itself. It should be noted that this omission is by no means a unique problem for literature focusing solely on algorithmic or AI harms. Bradley, for instance, argues that, despite the importance harm has across a range of domains, "almost nobody bothers to say what it is" [11:391]. Such an omission would not be a problem if harm were a simple, uncontroversial concept wherein relying on intuitive notions alone would suffice. Harm, however, is not simple, and there are often significant disagreements about what counts as harm [11:391]. As such, "it can be very difficult to get a clear picture of what exactly constitutes a harm" [67:8]. Moreover, relying on intuitive understandings of harm can render analyses themselves susceptible to biases, inconsistencies, and other inaccuracies. We agree with Birhane that intuitive understandings of harm are "poorly equipped to recognize injustice and oppression" [8:5] and make it far more likely that the harms experienced by certain groups will go undetected. AI researchers should be especially inclined to resist such a reliance given the complex and often unintuitive nature of artificial intelligence technologies [69,85].

3 DEFINING ALGORITHMIC HARM

There have been a variety of definitional accounts of what a harm consists of. These include non-comparative accounts, where one is harmed if they are made to be in a bad state [72] and event-based accounts where harm is viewed not as a state one is in but as a loss of certain basic goods [27]. However, scholarship that does define algorithmic harm typically relies on a well-known account by philosopher Joel Feinberg: harm as a setback to an interest. Feinberg offers a 'comparative' account of harm, wherein one is made worse off than they would have been had an activity not occurred [22]. Specifically, Feinberg characterizes harm as, "the thwarting, setting back, or defeating of an interest" [22:33]. Here interests are defined as something in which an individual has a stake; that is, in which the individual is "better or worse off depending on the condition of this thing" [quoted in 45:51]. According to Feinberg, humans can have interests in many things, including "physical integrity, intellectual acuity, a tolerable social and physical environment, and a certain amount of freedom from interest and coercion" [22:37]. Feinberg's account has appeared in multiple articles within the algorithmic harm literature. For instance, Solove and Citron draw on Feinberg's account to define harm to be "the impairment, or setback, of a person, entity or society's interests" [77:747]. Similarly, Smuha cites Feinberg in

defining harm as a form of setback to or thwarting of an interest [76:4], and Metcalf et al. characterize harm as setback to interests, where an interest is any outcome in which one has a stake [49:1455].

Though the understanding of harm as a setback to interests is perhaps the most common conceptualization of harm in philosophical literature, there are clear limitations to Feinberg's definition. First, an 'interest' is a philosophically complex concept and Feinberg himself spends considerable energy attempting to categorize the various dimensions of human "interest networks" [22:55]. Second, there is an inherent subjectivity in the holding of an interest; an individual may articulate a "stake" in any number of things for any number of reasons, and one could hold a stake in a thing or concept that might appear surprising or unreasonable to another. If said interest of such a person is set back, we might struggle to evaluate whether that individual has been truly harmed. Likewise, the opposite scenario is also plausible: an individual may not feel as though their interests have been set back, perhaps because of adaptive preferences, in situations where reasonable people would identify an obvious setback and therefore a clear harm. Such a comparative account of harm more generally has been subject to serious counterexamples [11,37,72]. For instance, a counterfactual comparative view such as Feinberg's faces what is referred to as the 'pre-emption problem'—if a given harm prevents a greater harm from occurring, counterfactual accounts do not register it as a harm, because one would not be worse off otherwise. Moreover, the extant philosophical literature on harm has itself been criticized for being 'highly idealized' and overly individualistic [17,50]. As several scholars have noted, considering harms beyond the individual level is especially relevant when thinking about digital technologies which have the potential to operate on broad or even global scales [8,65].

Nevertheless, Feinberg's account does supply crucial insights about harms. One is that there is a class of human experiences which do not necessarily rise to the level of harm, but which are still inimical to an individual's experience without being a "setback." Feinberg describes these experiences as "unhappy but not necessarily harmful" and classes them as a motley class of "hurts" and "offenses" [22:45–46]. These conditions are unpleasant but not *necessarily* harmful: in Feinberg's words, "an undesirable thing is harmful only when its presence is sufficient to impede an interest" [22:48]. Many of the examples of hurts and offenses Feinberg provides involve negative emotions or moments of subjective mental distress. Such phenomena often figure prominently in current debates regarding the impact of algorithmic technologies and AI systems in relation to social media effects [41,68,79], misinformation/disinformation[35,81,88], and Hoffmann's conception of dignitary harms and data violence [29,33]. We return to these phenomena and the circumstances under which they become harms—and wrongs—later in this paper.

Feinberg's account offers a second, even more crucial insight about harm—namely, that there is a conceptual difference between harms and wrongs such that not all harms are wrongs (and indeed, some wrongs may not even appear to be harms). In Feinberg's view, one wrongs another when one treats another unjustly, with "unjustly" understood as an action or omission that is "morally indefensible" [22:108]. In other words, wrongs necessarily hold moral considerations. In contrast, harms may include a normative dimension, but they need not do so. Adopting an example from [45], Person A would certainly be harmed if grievously injured by Person B; but if that injury came in response to Person A's own initial unprovoked attack on Person B, then Person A would not have also been wronged. This distinction between harms and wrongs has thus far been overlooked in the wider on algorithmic or AI-based harms, even by those who draw on Feinberg's account – for example, both [76] and [49] conflate the two concepts to define harm as *wrongful* setbacks to interests. In the following section, we argue that any suitable account of algorithmic harm should accommodate the distinction between harms and wrongs: both to avoid the various pitfalls of conflating harms and wrongs, but also to ground a more powerful and rigorous normative critique of the baleful effects of many algorithmic systems today.

4 ALGORITHMIC HARMS VS ALGORITHMIC WRONGS

Outside of the burgeoning literature on algorithmic harms, other philosophers have acknowledged the importance of distinguishing between harms and wrongs. Ben Bradley, for instance, argues that an adequate account of harm should avoid 'moralistic fallacies' and "not presuppose that harming is morally wrong, or involves vicious intent" [11:395]. For Bradley, this argument follows from a requirement for ontological neutrality, as harm can also be dealt by non-agents in cases where attributing wrongdoing would be inappropriate. For example, Bradley notes that "other sorts of events besides [human] actions are harmful too, like explosions and earthquakes" [11:394–395]. Though many might be harmed by such events, wrongdoing involves additional agential requirements that do not seem directly applicable in such contexts. There is a difference between harms caused by non-agents, harms caused by agents, and wrongs caused by agents. An earthquake itself cannot wrong a person; inadequate building codes or shoddy construction, in contrast, are harms that in such a case are also wrongs.

Another way to understand the significance of the distinction between harms and wrongs is by examining the issues which arise when the two concepts are conflated. Notably, most of the algorithmic harms considered and discussed in the current literature are harms that are also wrongs. Such a focus is understandable and appropriate given the greater moral significance of wrongful harms—the unjust character of certain algorithmic harms rightly makes them of primary concern. For example, the most pressing concerns regarding the societal impacts of AI systems come in contexts ranging from healthcare to policing, where setbacks to interests are clearly unjust in nature [14,21,57]. Similarly, concerns surrounding AI bias are often closely related to unjust or wrongful discrimination; privacy and security concerns often connect to rights violations; and issues of responsibility or accountability are particularly relevant in the case of wrongdoing. Recall, for instance, Hoffman's discussion of 'distributional harms' as injustices associated with the societal distribution of rights, responsibilities, and resources, and discursive data violence's deepening of oppressive structural positions [29,32]. The fact that harm is often conceptualized as an injustice, or synonymous with 'unjustified, unfair, or discriminatory effects' makes clear how pervasive the assumed

interchangeability of harms and wrongs is within the algorithmic harm literature [8,44,83].

The difference in moral status between harms and wrongs makes their conflation particularly problematic, however. For example, wronging typically involves violation or infringement – something that one has a claim against. What distinguishes harms and wrongs is what we are understood to 'owe' one another [3:80]. When A wrongs B, A violates B in a socially unwarranted manner that is often by way of a rights violation [3:81]. As a result, a claim of wronging presupposes different and often greater moral responsibilities or obligations than a claim of harming. Banja [3:80] observes that "someone's having experienced a harm doesn't necessarily mean that someone else is morally required to make a reparation." By focusing on harms that are also wrongs, but not acknowledging their status as wrongs, the current AI literature may inadvertently be *underemphasizing* the duties companies, regulators, and technologists have to address the adverse impacts of these technologies. Emphasizing the language of wrongdoing potentially offers additional normative support and guidance for the urgency of addressing the adverse impacts of AI systems.

By focusing implicitly on harms that are also wrongs, the current literature on algorithmic harms also overlooks two potentially notable categories of algorithmic impact: a) harms that are not wrongs, and b) wrongs that are not harms. These omissions narrow the scope of current analyses of algorithmic harms in important and potentially unhelpful ways. For example, consider a situation where an automated system used in a company's layoff decisions justifiably identifies a certain employee to be let go. Such a case would typically not be identified as an instance of algorithmic harm due to the typical conflation of harms and wrongs, but the layoff could significantly harm the employee without wronging them [3:80]. Likewise, some algorithmic decision-making systems in fields like medicine—such as those predicting the optimum radiation dose in cancer treatment—produce physical harm to the body without necessarily wronging the patient. In other medical contexts, algorithmic technologies for cancer screening that aim to predict risk inductively may also harm a patient without wronging them if they produce a false positive or a false negative result. Such harms may become wrongs if they are "unjustified;" in the medical context meaning going against clinical evidence and/or best practice. Yet even if such harms only remain harms, such cases of negative/adverse impacts require remedies. On the other hand, technologists and policymakers often need to translate broad but often diffuse concerns around algorithmic harm into actionable design decisions. Noting that some harms may not be wrongs in a normative sense signals such harms when they idiosyncratically subjective interests, are less urgent from a design/policy perspective.

Similarly, some wrongs created by algorithmic systems may be overlooked in current analyses if they are not also harms. Take, for instance, a modified example from [3]: a job seeker applies to both Company A and Company B. Though qualified for a job at either company, an algorithm used in Company A's hiring process unjustly rejects the job seeker's application on the grounds they are part of a protected class; Company B, however, provides a job offer which the job seeker gladly accepts. Even if the job seeker actively disliked Company A and would not have accepted a job if it was offered, they are still wronged—but not necessarily harmed—by Company A's hiring algorithm. As Banja elaborates, "A's withholding the job offer was not ethically or legally justified because it was a rights violation surrounding discrimination and was therefore an instance of wronging." Yet the job seeker, as they defined their own interests, would not have chosen to work for Company A in any case [3:81]. An even more salient example comes from Obermeyer et al.'s [56] well-known analysis of a commercial prediction algorithm used to assign extra care in a large hospital system. The algorithm was found to be biased against Black patients because it misidentified cost of care with severity of illness. The white patients who benefited from this misallocation of care were not harmed—quite the contrary. However, they along with the Black patients involved, were wronged insomuch as they were treated unjustly within the scope of the system (indeed, even Black patients whose care was not directly impacted by the algorithm were potentially also wronged insofar as they relied on a health care system that discriminated against them). However, this example also illustrates Feinberg's observation that "there *can* be harms that are not wrongs *on balance*, but there are few wrongs that are not *to some extent* harms" [22:35]: in the case studied by Obermeyer et al., the wrong inherent in the system's bias produced strong harms for Black patients, while the wrong was abstract and unfelt by white patients.

Distinguishing between harms and wrongs is necessary if we are to accurately track the effects of discriminatory biases present in certain algorithms, regardless of their material impact on individual interests. It is imperative that scholars and policymakers employ the appropriate conceptual tools to help provide normative clarity on how to respond to those far-reaching impacts: and one of those tools is recognizing the distinction between harms and normatively weighted wrongs, as well as the impacts of "rare nonharmful wrongs and common nonwrongful harms" [22:36]. We therefore argue that there are especially good reasons to be attuned to the conceptual difference between harms and wrongs in the context of artificial intelligence systems. The uptake of these systems in a growing number of domains—from healthcare to education to law (and more)—means that for most of us, AI pervades our everyday lives. As noted above, the stakes are often high in the contexts of the systems' use. AI technologies can encroach or invade our fundamental interests like privacy, freedom, and autonomy, and can reinforce and amplify instances of structural injustice, as well as deleterious networks of power and oppression. Ultimately, the scope of these algorithmic systems means that their use will have pervasive impacts [48,52]. Understanding the harms/wrongs distinction is one way that we can begin to unpack and address these impacts.

5 UPHOLDING THE HARMS/WRONGS DISTINCTION IN ALGORITHMIC CONTEXTS

What practical work does the distinction between algorithmic harms and algorithmic wrongs do in identifying and remedying the adverse impacts of algorithmic systems? First and foremost, we argue that the process of distinguishing between harms and wrongs provides a mechanism for value prioritization in design

and engineering contexts. Creating technical objects always entails making choices that represent particular values and interests: as Flanagan and Nissenbaum observe, "any functioning artifact is the product of interacting (and sometimes conflicting) constraints, including physical, economic, and functional constraints" [24:108]. Grappling with this array of conflicting interests makes it difficult to avoid harm entirely as an outcome of the development of an algorithmic system, again understanding harms specifically as setbacks to some interest or another [24:112,73]. When making such a tradeoff, we argue that designers and developers should first prioritize avoiding *wrongful harms*, and then *systematic harms* that are adequately foreseeable. Though these recommendations may seem commonsensical, they are a non-trivial problem for technologists in the throes of the development process [34,61]. Van den Hoven, Lockhorst, and van de Poel observe that engineers bear a particular responsibility, "to prevent situations which are morally dilemmatic, and which must inevitably lead to suboptimal solutions or compromises and trade-offs from a moral point of view" [34:144]. Making a clear distinction regarding which interests are normatively charged and which are not is a crucial first step in implementing such a principle—one in which design methods geared towards exploring competing values in technical design are often helpful [24,25].

Some developers might protest that it is difficult to know how and when competing interests or values tradeoffs lead to systematic harms and/or become wrongs. We sympathize but emphasize that close adherence to a harm/wrongs distinction actually makes such judgements more straightforward. If an AI system can be reasonably anticipated to produce an algorithmic wrong before it is developed or deployed, it should not be created in the first place (as a reminder, a wrong is when one is treated unjustly, with "unjustly" understood as an action or omission that is "morally indefensible" [22:108]). To reiterate a key point, "not all problems can or should be solved with [algorithmic] technology" [70:13], not least those whereby such technologies exacerbate or reinforce existing wrongs. But even when designers do not anticipate a technical system will cause wrongdoing, they can be assured that it will produce some range of harms. As such, even after attending for the potential to do wrong, designers should still work to minimize systematic harms and interest and value conflicts in their technologies through mechanisms such as creative redesigns to dissolve or mitigate such antipathies [24:109–113]. The case of algorithmic systems used in medical contexts is illustrative: it may well be a wrong if a designer of such systems does not make every attempt to minimize reasonable harm to patients, but even above this standard, designers and technologists are often impelled to shift the standard of "reasonableness" towards less systematic harm.

Inevitably, an AI system will prioritize some interests over others, and thus potentially produce algorithmic harms. As noted above, non-wrongful harms can be significant in themselves and should be avoided if possible. Moreover, tracking the responsibilities and impacts of potential harms—especially in highly formalized instances like the medical context—is necessary, in part because a harm as *cause* may give rise to a new wrong as *outcome*. To accurately track this distinction between harms and wrongs, we argue further that certain requirements need to be met within both the scholarship and design practice dedicated to these topics. For the purposes of this paper, we focus on two examples. First, we argue that it is necessary to understand harms and wrongs as contextual. Second, we suggest that tracking this distinction requires engaging diverse communities and stakeholders. This second requirement flows from the first; because harms are contextual and socially embedded, we need to consult with a wide range of stakeholders to appropriately conceptualize both AI's harmful and wrongful impacts.

The idea that harm is context-dependent has often been met with resistance in the philosophical literature. For instance, Johannsson and Risberg [37:23] argue that there is 'no immediate reason' to think that whether someone is harmed depends on context.[1] In some cases, however, identifying a harm does seem to require contextual analysis. For example, an AI system used for medical diagnosis which does not provide patients with an explanation for its results may not constitute a harm to everyone [42]. However, for a patient who has experienced the prior wrong of discriminatory treatment in health care, a lack of explanation could reinforce feelings of marginalization or lead to non-compliance with treatment, causing significant harm [28:4]. Similarly, harm theorists have repeatedly stressed that an adequate theory of harm "needs to enable us to measure the severity of harm" [82]. For instance, though privacy breaches may constitute a harm no matter who you are, they can engender particularly severe harms depending on identity and circumstance: they could, for instance, "cause [individuals] with certain stigmatized illnesses to be alienated from their communities [,] reduce a person's opportunities for employment, [or] even lead to unwarranted increases in health insurance costs" [78:89] As these examples above suggest, social and systemic factors greatly impact the severity of a given individual harm.

Moreover, though it may be possible to identify some harms without considering wider contextual factors, distinguishing between harms and wrongs *does* seem to hinge on context. An assessment of wronging or wrongdoing typically entails an assessment of how a harm (or non-harm wrong) came about. As Duff explains, one thing that distinguishes harms from wrongs is the matter of how they are caused. "One whose welfare interests are set back by a wrongful human action might suffer just the same harm as one whose welfare interests are set back by natural causes," Duff notes; "What makes the former's harm a [wrong] is the additional fact that the harm was caused by conduct that wronged him" [20:18]. To determine whether conduct is wrongful, an adjudicator needs to consider a variety of contextual factors ranging from agents' intentions and knowledge to the social norms governing unwarranted action and what we are understood to 'owe' one another. A general example of why this is so can be taken from [74]: imagine A accidentally cut her finger with a knife while cooking. At first glance, it appears that A has suffered a harm, but not a wrong. However, if the accident occurred within a highly unsafe work environment, it seems far more likely that the harm A experiences also constitutes a wrong [74:126]. Likewise, offences, or "conditions are unpleasant but not necessarily harmful," are only harmful if the offence is reasonable: hence the importance of context to distinguishing between harms

[1] To demonstrate why this is the case, the authors compare the sentence 'e harms S' to a paradigmatically context-sensitive sentence of 'S is tall' (p.23). The theorists argue that 'S is tall' can be true in one context (e.g. when S is compared to her colleagues), but false in another (e.g. when they are compared to basketball players), but that 'e harms S' does not behave in this way (p.24).

and wrongs. The reasonableness of taking offence depends on several factors, including the temporal and spatial durability of what has caused the offence (e.g., I may be offended by the smell of dog feces on the ground, but am not harmed if I have the ability to walk away from it), the extent to which I have sought out a situation knowing in advance I will find it unpleasant, and/or the extent to which my sense of unpleasantness derives from visceral versus reflective responses.

Distinguishing between the harmful and wrongful impacts of algorithmic systems also necessitates analysis of the broader social context from which these systems emerge and to which they are deployed [8]. For example, consider the issue of 'proxy attributes' —seemingly innocuous data points that correlate with socially-sensitive attributes which serve as proxies for the socially-sensitive attributes themselves [36,38,54]. To understand why systems harboring biases associated with a particular zip code are not just harmful but wrongful, analysts must consider the social contexts in which this data is embedded, acknowledging, for instance, the relationship between neighborhood demographics and race as the product of historical and ongoing patterns of racial oppression [7]. The need to consider context—especially social context—when analyzing the effects of these technologies is a touchstone of critical scholarship on algorithmic systems [6,9,28,55,85,89]. In criticizing overly technical solutions to algorithmic harm, Birhane argues that "a fundamentally equitable path must examine the wider picture, such as unquestioned or intuitive assumptions in datasets, current and historical injustices, and power asymmetries" [8:1]. Himmelreich and Lim similarly stress the need to analyze social and structural contexts, claiming that "in short, social structures explain the patterns of behavior and phenomena that data 'represent,' and social structures condition practices that generate these data" [28:7]. These and many other calls for analysis of algorithmic systems as socially embedded underscore the point that, contra the extant philosophical literature, harms and wrongs are highly contextual concepts.

Second, we argue there is an urgent need to engage with a multitude and a diversity of stakeholders and community members on the harmful and wrongful impacts of AI systems. It is often easiest to viscerally comprehend harms and wrongs we ourselves have experienced, and those who are affected by new technologies are often epistemically best positioned to understand the nature of those impacts [5,75]. Work in feminist epistemology has long recognized that individuals can hold valuable and epistemically consequential insights in virtue of their social positionality [87], including insights about when and whether they and/or their community have been harmed, wronged, or both. Recent work on algorithmic harm has increasingly emphasized the value of acknowledging the capaciousness and multifariousness of expertise in evaluating the significance of new AI systems in this way. Moss et al., for example, write that "understanding algorithmic harms requires a broader community of experts: community advocates, labor organizers, critical scholars, public interest technologists, policy makers, and the third-party auditors who have been slowly developing the tools for anticipating algorithmic harms" [52]. Similarly, Metcalf et al. note that expertise is "not limited to professional capacities," and that "Individuals and communities affected by algorithmic systems are often the foremost experts in the potential harms they regularly encounter, as well as the strategies they have developed to minimize or avoid such harms" [48]. This argument has a direct impact on policymaking, regulation, and standard setting in the context of AI and other algorithmic systems [80]. To understand when and how harms represent wrongs, such diversity of evidence is crucial.

One example is the case of vaccine passports: a new health technology that allows individuals to carry easily accessible documentation of their vaccination status on their mobile devices [53]. Though such an innovation might seem convenient and harmless, Mukogosi points out that certain groups, such as the BIPOC community, may be distrustful towards medical professionals thanks to a long history of medical violence directed towards them. This distrust might lead to a justified unwillingness to undergo vaccination, and the oft misunderstood and underrecognized history of this problem can result in discrimination if vaccine passports are required for entry in certain locations. If a designer has never had reason to distrust institutional healthcare, then they are unlikely to consider the inequitable and potentially harmful impacts of introducing health passports as a new technology for everyday use.

A diverse set of perspectives is vital to grasp the full impact of AI technologies in the many contexts in which these technologies are being deployed, and to collectively address them. This second criteria for respecting Feinberg's harms/wrongs distinction flows from the first; it is because harms are contextual and relational in nature, policymakers need to engage with those who are subject to them to understand how societies should respond appropriately.

6 CONCLUSION

We have argued that, despite widespread concerns surrounding the harms produced algorithmic systems, the concept of harm itself remains under-analyzed in the relevant literature. As a result, most articles fail to explicitly define harm, rendering analyses susceptible to biases and other inaccuracies. One notable exception to this trend is work on algorithmic harm that rely on a definition of harm developed by Joel Feinberg, wherein harms are conceptualized as setbacks to interests [49,76]. Though Feinberg's account of harm has several shortcomings, it nonetheless points us in the direction of an improved account of algorithmic harm. We argue that Feinberg's distinction between harms and wrongs, which has thus far been overlooked in the current scholarship on algorithmic harm, is a necessary feature of any full account of these technologies' impacts. The conflation of harms and wrongs when analyzing the impacts of AI has led to weakened claims of obligation in response to the wrongful harms of AI, and the failure to recognize the wrongless harms as well as the harmless wrongs produced by these technologies. In advocating for the distinction between harms and wrongs, we have argued both for the need to understand both harms and wrongs as contextual, socially embedded concepts, and have suggested that tracking this distinction requires engagement with a diverse range of communities and stakeholders. In so doing, our hope is to lay the groundwork for the development of a suitable theory of algorithmic harm that tracks the conceptual and moral differences between harms and wrongs, and one which has impact on both the design of and policymaking around algorithmic systems of all kinds.

ACKNOWLEDGMENTS

The authors would like to acknowledge the feedback and assistance of numerous individuals, including Carolyn McLeod, Joanna Redden, Dan Lizotte, Benjamin Chin-Yee, Emily Cichocki, Danica Pawlik-Potts, Jason Millar, Heather Stewart, Jacquelyn Burkell, Maxwell Smith, Kira Lussier, Hale Doguoglu, and other participants in the symposium on "(Dis)Trust and AI: Perspectives from Across Disciplines and Sectors" held at Western University in October 2023; Jacob Metcalf, Emanuel Moss, Jay Shaw, Andrew Buzzell, Joseph Donia, and the participants of "The Social Life of Algorithmic Harms" workshop hosted by the Data & Society Research Institute in March of 2022; and funding support from the Social Sciences and Humanities Research Council of Canada (SSSHRC) (grant # 430-2021-01041), the Faculty of Information and Media Studies (FIMS) and Faculty of Arts and Humanities at Western University, and the university's Interdisciplinary Development Initiatives, Undergraduate Summer Research Internships, and Radboud-Western Collaboration Funds.

REFERENCES

[1.] Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, *et al.* 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine* 4, 1: 65.

[2.] Julia Angwin. 2010. The What They Know Series. Retrieved from http://juliaangwin.com/the-what-they-know-series/.

[3.] John Banja. 2001. When Harms Become Wrongs. *Journal of Disability Policy Studies* 12, 2: 79–86.

[4.] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3: 671–732.

[5.] Ruha Benjamin. 2016. Informed Refusal. *Science, Technology, & Human Values* 41, 6: 967–990.

[6.] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464: 421–422.

[7.] Ruha Benjamin. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

[8.] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2: 100205.

[9.] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2: 100205.

[10.] danah boyd and Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15, 5: 662–679.

[11.] Ben Bradley. 2012. Doing Away with Harm. *Philosophy and Phenomenological Research* 85, 2: 309–412.

[12.] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. 1–15.

[13.] Alan Chan, Rebecca Salganik, Alva Markelius, *et al.* 2023. Harms from Increasingly Agentic Algorithmic Systems. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 651–666.

[14.] Benjamin Chin-Yee and Ross Upshur. 2019. Three Problems with Big Data and Artificial Intelligence in Medicine. *Perspectives in Biology and Medicine* 62, 2: 237–256.

[15.] Mark Coeckelbergh. 2020. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics* 26, 4: 2051–2068.

[16.] Nick Couldry and Ulises A Mejias. 2019. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media* 20, 4: 336–349.

[17.] Shannon Dea. 2020. Toward a Philosophy of Harm Reduction. *Health Care Analysis* 28, 4: 302–313.

[18.] Lina Dencik, Joanna Redden, Arne Hintz, and Harry Warne. 2019. The 'golden view': data-driven governance in the scoring society. *Internet Policy Review* 8, 2: 1–24.

[19.] Nathalie DiBerardino and Luke Stark. 2023. (Anti)-Intentional Harms: The Conceptual Pitfalls of Emotion AI in Education. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 1386–1395.

[20.] R A Duff. 2001. Harms and Wrongs. *Buffalo Criminal Law Review* 5, 1: 13–45.

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

[22.] Joel Feinberg. 1987. *The Moral Limits of the Criminal Law Volume 1: Harm to Others*. Oxford University Press, New York, and Oxford.

[23.] Benjamin Fish and Luke Stark. 2022. It's Not Fairness, and It's Not Fair: The Failure of Distributional Equality and the Promise of Relational Equality in Complete-Information Hiring Games. *Equity and Access in Algorithms, Mechanisms, and Optimization*: 1–15.

[24.] Mary Flanagan and Helen Nissenbaum. 2014. *Values at Play in Digital Games*. The MIT Press.

[25.] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design*. The MIT Press.

[26.] Maya Indira Ganesh and Emanuel Moss. 2022. Resistance and refusal to algorithmic harms: Varieties of 'knowledge projects.' *Media International Australia* 183, 1: 90–106.

[27.] Matthew Hanser. 2008. The Metaphysics of Harm. *Philosophy and Phenomenological Research* 77, 22: 421–450.

[28.] Johannes Himmelreich and Désirée Lim. 2022. The Oxford Handbook of AI Governance. .

[29.] Anna Lauren Hoffmann. 2016. Beyond distributions and primary goods: Assessing applications of rawls in information science and technology literature since 1990. *Journal of the Association for Information Science and Technology* 68, 7: 1601–1618.

[30.] Anna Lauren Hoffmann. 2018. Data Violence and How Bad Engineering Choices Can Damage Society. Retrieved from https://medium.com/s/story/data-violence-and-how-bad-engineering-choices-can-damage-society-39e44150e1d4.

[31.] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7: 900–915.

[32.] Anna Lauren Hoffmann. 2020. Terms of inclusion: Data, discourse, violence. *New Media & Society* 81, 2: 146144482095872–18.

[33.] Anna Lauren Hoffmann. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society* 23, 12: 3539–3556.

[34.] Jeroen van den Hoven, Gert-Jan Lokhorst, and Ibo van de Poel. 2011. Engineering and the Problem of Moral Overload. *Science and Engineering Ethics* 18, 1: 143–155.

[35.] Caroline Jack. 2017. Lexicon of Lies: Terms for Problematic Information. 1–22.

[36.] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 375–385.

[37.] Jens Johansson and Olle Risberg. 2023. A Simple Analysis of Harm. *Ergo an Open Access Journal of Philosophy* 9, 19.

[38.] Gabbrielle M. Johnson. 2021. Algorithmic bias: on the implicit biases of social technology. *Synthese* 198, 10: 9941–9961.

[39.] Oscar H Gandy Jr. 1996. Coming to Terms with the Panoptic Sort. In D. Lyon and E. Zureik, eds., 133–155.

[40.] Ian Kerr. 2010. Digital Locks and the Automation of Virtue. In 247–303.

[41.] Daniel Kreiss and Shannon C McGregor. 2023. A review and provocation: On polarization and platforms. *New Media & Society*: 146144482311618.

[42.] Alex John London. 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report* 49, 1: 15–21.

[43.] David Lyon. 2007. Data, Discrimination, Dignity. In 179–197.

[44.] Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. 2021. Algorithmic pollution: Making the invisible visible. *Journal of Information Technology* 36, 4: 391–408.

[45.] Carolyn McLeod. 2020. *Conscience in Reproductive Health Care*. Oxford University Press, New York, and Oxford.

[46.] Dan McQuillan. 2022. *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Bristol University Press, Bristol, UK.

[47.] Andrew McStay. 2020. Emotional AI, soft biometrics, and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society* 7, 1: 2053951720904386.

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

[49.] Jacob Metcalf, Ranjit Singh, Emanuel Moss, Emnet Tafesse, and Elizabeth Anne Watkins. 2023. Taking Algorithms to Courts: A Relational Approach to Algorithmic Accountability. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 1450–1462.

[50.] Sarah Clark Miller. 2022. Toward a relational theory of harm: on the ethical implications of childhood psychological abuse. *Journal of Global Ethics* 18, 1: 15–31.

[51.] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* 22, 4: 16–28.

[52.] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*. Data & Society Research Institute.

[53.] J. Mukogosi. 2021. Vaccine Passports and Health Racism. *Medium*. Retrieved January 3, 2024 from https://points.datasociety.net/vaccine-passports-and-health-racism-7e494e29bd9b.

[54.] Dylan Mulvin. 2021. *Proxies: The Cultural Work of Standing In*. The MIT Press, Cambridge, MA.

[55.] Helen Nissenbaum. 2015. Respecting Context to Protect Privacy: Why Meaning Matters. *Science and Engineering Ethics* 109, 4: 1–22.
[56.] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. 89–89.
[57.] Cathy O'Neil. 2017. *Weapons of Math Destruction*. Broadway Books.
[58.] Mimi Onuoha. 2018. Notes on Algorithmic Violence. Retrieved January 1, 2024 from https://github.com/MimiOnuoha/On-Algorithmic-Violence.
[59.] Frank Pasquale. 2015. *The Black Box Society*. Harvard University Press.
[60.] Simon A. Pemberton. 2016. *Harmful Societies: Understanding Social Harm*. Policy Press, Bristol, UK.
[61.] Ibo van de Poel and Peter-Paul Verbeek. 2006. Editorial: Ethics and Engineering Design. *Science, Technology, & Human Values* 31, 3: 223–236.
[62.] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. 2022. A Human Rights-Based Approach to Responsible AI. *arXiv*.
[63.] Michael Rabenberg. 2014. Harm. *Journal of Ethics & Social Philosophy* 8, 3: [viii]-31.
[64.] Bogdana Rakova and Roel Dobbe. 2023. Algorithms as Social-Ecological-Technological Systems: An Environmental Justice Lens on Algorithmic Audits. *2023 ACM Conference on Fairness, Accountability, and Transparency*: 491–491.
[65.] Joanna Redden, Jessica Brand, and Vanesa Terzieva. 2020. Data Harm Record. Retrieved January 16, 2024 from https://datajusticelab.org/data-harm-record/.
[66.] Joanna Redden, Lina Dencik, and Harry Warne. 2020. Datafied child welfare services: unpacking politics, economics, and power. *Policy Studies* 41, 5: 507–526.
[67.] Diane Riley and Pat O'Hare. 2000. Harm Reduction: History, Definition, and Practice. In J.A. Inciardi and L.D. Harrison, eds., *Harm Reduction: National and International Perspectives*. SAGE Publications Inc., Thousand Oaks, CA.
[68.] Nick Seaver. 2019. Captivating algorithms: Recommender systems as traps. *Journal of Material Culture* 24, 4: 421–436.
[69.] Andrew D. Selbst and Solon Barocas. 2018. The Intuitive Appeal of Explainable Machines. *Fordham Law Review* 87, 3: 1085–1139.
[70.] Andrew D. Selbst, danah boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. 59–68.
[71.] Renee Shelby, Shalaleh Rismani, Kathryn Henne, *et al.* 2022. Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction. *arXiv*.
[72.] Seana Shiffrin. 2012. Harm and Its Moral Significance. *Legal Theory* 18: 357–398.
[73.] Katie Shilton. 2018. Engaging Values Despite Neutrality. *Science, Technology, & Human Values* 43, 2: 247–269.
[74.] Thomas W. Simon. 1995. Group Harm. *Journal of Social Philosophy* 26, 3: 123–137.
[75.] Mona Sloan, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a Design Fix for Machine Learning. 1–7.
[76.] Nathalie Smuha. 2021. Beyond the individual: governing AI's societal harm. *Internet Policy Review* 10, 3.
[77.] Daniel J. Solove and Danielle Keats Citron. 2016. Risk and Anxiety: A Theory of Data Breach Harms. *SSRN Electronic Journal*.
[78.] Robert Sparrow and Joshua Hatherly. 2019. The Promise and Perils of AI in Medicine. *International Journal of Chinese & Comparative Philosophy of Medicine* 17, 2: 79–109.
[79.] Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science* 48, 2: 204–231.
[80.] Luke Stark, Daniel Greene, and Anna Lauren Hoffmann. 2021. Critical Perspectives on Governance Mechanisms for AI/ML Systems. In J. Roberge and F. McKelvey, eds., *The Cultural Life of Machine Learning*. Palgrave Macmillan, 257–280.
[81.] Francesca B. Tripodi, Lauren C. Garcia, and Alice E. Marwick. 2023. 'Do your own research': affordance activation and disinformation spread. *Information, Communication & Society* ahead-of-print, ahead-of-print: 1–17.
[82.] Charlotte Franziska Unruh. 2023. A Hybrid Account of Harm. *Australasian Journal of Philosophy* 101, 4: 890–903.
[83.] Shannon Vallor. 2015. Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology* 28, 1: 107–124.
[84.] Carissa Véliz. 2022. The Oxford Handbook of AI Governance. .
[85.] Jess Whittlestone and Sam Clarke. 2022. The Oxford Handbook of AI Governance. *arXiv*.
[86.] Pak-Hang Wong. 2020. Democratizing Algorithmic Fairness. *Philosophy & Technology* 33, 2: 225–244.
[87.] Alison Wylie. 2003. Why Standpoint Matters. In R. Figueroa and S. Harding, eds., *Science and Other Cultures: Issues in Philosophies of Science and Technology*. Routledge, New York, 26–48.
[88.] Jason C. Young. 2021. Disinformation as the weaponization of cruel optimism: A critical intervention in misinformation studies. *Emotion, Space and Society* 38: 100757.
[89.] Matthew Zook, Solon Barocas, danah boyd, *et al.* 2017. Ten simple rules for responsible big data research. *PLoS computational biology* 13, 3: e1005399-10.