Algorithmic Accountability and Digital Justice: A Critical Assessment of Technical and Sociotechnical Approaches

Howard Rosenbaum
Indiana University, Bloomington, Indiana, USA
hrosenba@indiana.edu

Pnina Fichman
Indiana University, Bloomington, Indiana, USA
fichman@indiana.edu

ABSTRACT

The concept of digital justice is intended to open up discourse about strategies for bringing relief to those who believe they have been discriminated against or harmed by algorithmic decision making. Digital justice has depended on algorithmic accountability, a means by which entities can be held accountable for the consequences of algorithmic decision making. This paper critically examines the concept of algorithmic accountability to assess its utility as a ground for digital justice and argues that it is fraught with difficulties. After discussing digital justice and algorithmic discrimination, algorithmic accountability is decomposed into two types, technical and sociotechnical. These approaches are critically assessed and a cautionary note is struck about the difficulty of enacting algorithmic accountability. If this argument is persuasive, it implies that the concept of digital justice also has difficulties. The paper concludes with suggestions for moving forward that do not use either version of algorithmic accountability.

KEYWORDS

Algorithmic Accountability; Digital Justice

ASIS&T THESAURUS

Information Science; Data Science

INTRODUCTION

Algorithms, as key components of algorithmic assemblages, are having an increasingly significant impact on our lives. From activities as mundane as online searching, using GPS directions, and seeking online recommendation to those as complex as making loan decisions, hiring, and making parole decisions, algorithmically mediated interactions are shaping our social and work lives in ways that are just beginning to come into focus. As Martin (2018, p. 1) explains, "[a]lgorithms can determine whether someone is hired, promoted, offered a loan, or provided housing as well as determine which political ads and news articles consumers see." Striphas (2015, p. 395) observes that "over the last 30 years or so, human beings have been delegating the work of culture – the sorting, classifying and hierarchizing of people, places, objects and ideas – increasingly to computational processes." This has, he argues (2015, 396) led to "the enfolding of human thought, conduct, organization and expression into the logic of big data and large-scale computation." Willson (2016, p. 140) suggests that this trend is important because "the ways algorithms are designed and implemented (and their resultant outcomes) help to influence the ways we conduct our friendships (Bucher, 2012), shape our identities (Cheney-Lippold, 2011) and navigate our lives more generally (Beer, 2009)." Put succinctly, "[a]lgorithms are not immaterial formulae, but practical expressions that that affect the phenomenal word of people" (Klett, 2016, p. 112).

Scholars working in such domains as critical data studies, legal scholarship, computer science, information science, anthropology, and sociology have begun to grapple with the complexities of the sociotechnical impacts of the integration and routinization of algorithmic assemblages into social and organizational life. Scholars critical of the potential for discriminatory outcomes of algorithmically-driven decision making in such domains as banking, insurance, college admissions, and criminal justice have begun to advocate for ways to hold the organizations that own and/or control algorithms accountable for these outcomes. In this move toward what some are calling "digital justice," (Couldry, Gray, & Gillespie, 2013; Jentile & Lawrence, 2016; Sabelli & Tallacchini, 2018; Taylor, 2017), the concept of algorithmic accountability has been introduced as a key component in the attempts to redress the problems created by reliance on algorithmic assemblages for decision making and support.

In this paper we critically examine the concept of algorithmic accountability to assess its utility as a ground for an approach to digital justice. After briefly describing the concepts of digital justice and algorithmic discrimination, algorithmic accountability is decomposed into two types, technical and sociotechnical. We provide a critical assessment of these approaches that strikes a cautionary note about the difficulty of enacting algorithmic accountability and the subsequent consequences for possibility of digital justice for those harmed by algorithmic decision making. The paper concludes with suggestions for moving forward that do not rely on either of these approaches.

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19–23 October, 2019

DOI: 10.1002/pra2.00019

DIGITAL JUSTICE

The concept of digital justice is, at the moment, variegated according to the lens that is used to examine it; data justice can be confrontational, distributive, or protective. In its confrontational form, it is a challenge to the uses of data to support and reinforce existing power asymmetries between those who gather and analyze the data and those who generated the data (Barnett, Koshiyama, & Treleaven, 2017, p. 1; Diakopoulos, 2016, p. 58; Martin, 2018, p. 3; McCarthy, 2016, p. 1133; Prins, 2018). According to this critique "data systems tend to have a disciplinary function because the way data are collected and structured constitutes a form of normative coercion" (Taylor, 2017, p. 6). Digital justice can also be oriented toward distributive uses of information with the goal of foregrounding groups disenfranchised by the effects of the big data divide (Barocas, Bradley, Honavar, & Provost, 2017, p. 3; McCarthy, 2016, p. 1131) that characterize the modern information environment. Finally, it can be protective, with the goal of mitigating the power of the surveillance state, which uses its ability to collect vast amounts of data about citizens (Fink, 2017, p. 13). This approach to digital justice emphasizes "resistance to government surveillance based on principles of social justice" (Taylor, 2017, p. 7).

One characteristic that is common to these approaches is that digital justice is not based on a foundation of individual rights because "data injustice increasingly tends to occur on the collective level" (Taylor, 2017, p. 8). Big data analytics aggregate data and generate insights about groups and populations meaning that digital justice is better seen as a structural concept that involves concern for preserving the privacy of groups in part by enhancing their ability to control their engagement with data and data organizations which, in turn, extends their capacity to control their data visibility, and their rights to non-discriminatory interactions with institutions that gather big data and make use of big data analytics. The positive dimension of digital justice involves a drive to develop ways to protect and enhance people's involvement with big data. This may involve a move toward algorithmic regulation (Mcquillan, 2017, p. 568). The negative dimension involves developing ways to redress people's grievances when they believe that they have been wronged by institutional algorithmically-mediated decision making. Because the problem of algorithmic discrimination is central to digital justice, the move to hold institutions responsible for these actions currently relies on a concept of algorithmic accountability.

ALGORITHMIC DISCRIMINATION

Algorithms can be used to enhance or hinder social progress and digital justice. The social life of algorithms is shaped in part by a pervasive technological frame that casts them as objective and neutral, meaning that their outputs are free from bias (Binns, 2018, p. 546; Iliadis, 2018, p. 3). This frame is supported by stakeholders with vested interests in preserving the status of algorithmic assemblages as neutral because it benefits them economically, legitimates them socially, and preserves power asymmetries. The frame masks what Martin (2018, p. 2) describes as a false tension - "algorithms as objective, neutral blank slates versus deterministic, autonomous agents;" when framed in this way, the discourse "absolves firms of responsibility for the development or use of algorithms" because the clean code is used in messy contexts with contaminated data by people who do not know how to use them. Those using algorithmic tools respond by claiming that the algorithms are too complex and opaque, so their responsibility for outcomes is mitigated.

Behind the veil of objectivity, critics argue that algorithms are value laden, have moral consequences and can reinforce or challenge power asymmetries (Ananny & Crawford, 2018, p. 978; Mccann, Hall, & Warin, 2018, p. 14). More specifically, they can support or undermine ethical qualities of the domain in which it is intended to operate this is due, in large part, to the decisions made by developers as they create the algorithms. The types of decisions that are relevant here are those that determine the roles the end users will play in the algorithmically-mediated decision; in simple terms, the decision can be left to the algorithm (with no human intervention), to the end users (with no algorithmic intervention), or to some combination of both (Martin, 2018, p. 2). In making these decisions, designers and developers are also are delegating accountability in the decision-making process.

Another important type of decision made by designers and developers is about the attributes of the data set that, in their opinions, are critical inputs to be taken into account as the algorithm generates its output. Two assumptions underlie big data collection that affect the training sets used to train algorithms. The first is that the training set is an accurate representation of the future populations to which the algorithm will be applied. The second is that the sample used in the training set is a good representation of the individuals constituting that population. However, "these assumptions are susceptible to error and bias, although that is precisely what they are intended to negate" (Haarkens, 2018, p. 22). Some subset of these attributes will be appropriate to the decision context while others may not; an example is the inclusion of attributes that are proxies for race in the algorithms used in the COMPAS system, which "wrongly labeled defendants as 'future criminals' when they did not commit a crime at twice the rate for black defendants as white defendants" (Martin, 2018, p. 4). In a sense, the developers are inscribing into the algorithms their assumptions about how the artifact will fit into the social context for which it is intended (Neyland, 2019, p. 32).

Just as there are legal prohibitions against discrimination by humans, there are restrictions against discrimination by algorithms (Kleinberg, Ludwig, Mullainathan, & Sunstein, 2019, p. 2). The challenge is to provide a level of proof that stands up to legal scrutiny. They argue against opening the algorithmic black box because it is a "mathematical impossibility" to understand what an algorithm will do by reading its code; the data it uses must be examined and the algorithm must be run and its outcomes observed (Kleinberg et al., 2019, p. 2). This allows an analyst to determine whether some input into the algorithm led to a discriminatory outcome or whether there is some other externality responsible for the observed disparity. There must be documentation of the decisions made during the development and training process, for example, the predictive goal used to train the algorithm affects the analysis of the training at a and the resulting outcome. The implication of this approach is that it is more important to regulate the person developing the algorithm than the algorithm itself.

Kleinberg et al. (2019, p. 21) describe four types of algorithmic discrimination. An algorithm may produce outcomes that lead to disparate treatment, predicting on the basis of gender or race. It may result in disparate impact, using a predictor variable that disadvantages a vulnerable population. It may use a predictor variable that is a result of prior discrimination, such as a credit score. Finally, it may produce a result that people find disturbing, such as giving preferential treatment to one gender over another; of the four, they point out that this last type of discrimination is not covered by law.

ALGORITHMIC ACCOUNTABILITY

A widespread and powerful technological frame has developed about algorithms and algorithmic assemblages over the last decade. It has taken the form of a decidedly utopian discourse about algorithms (and, by implication, the algorithmic assemblages in which they are embedded) that portrays them as objective, rational, and authoritative (Dourish, 2016; Gillespie, 2014; Lee, 2018). This technological frame is based, in part, on the way in which algorithms are thought of in their primary domain of origin, computer science. They are seen as code that becomes part of a software program that, when activated, can carry out sorting, optimizing, and prioritizing tasks quickly, efficiently, and without human intercession. Given this, algorithms are typically seen by computer scientists as tools that can be used to provide control plus logic (Lustig & Nardi, 2015, p. 744) in software. According to Gillespie (2014, p. 4):

> More than anything, algorithms are designed to be and prized for being functionally automatic, to act when triggered without any regular human intervention or oversight.

Despite the optimistic discourse about the power of algorithms to improve lives, Rainie and Anderson (2017, p. 1) report that "experts worry they can also put too much control in the hands of corporations and governments, perpetuate bias, create filter bubbles, cut choices, creativity and serendipity, and could result in greater unemployment." McCann et al. (2018, p. 14) argue that

> Although algorithms held out the promise of a more neutral decision making process, in reality it is more accurate to think of them as 'an opinion embedded in mathematics/

Willson (2016; p.145) foreshadowed this insight, claiming that the "combination of delegated everyday practices and algorithmic functions within social, cultural and political systems inescapably results in biases being enacted." Diakopoulos (2014, p. 2) states flatly that "[w]e're living in a world now where algorithms adjudicate more and more consequential decisions in our lives." In the domain of search engines, Willson (2016, p. 143) notes that "Google's Panda, Penguin and Hummingbird algorithms and recurrent updates are other examples where changes are made in order to encourage some outcomes, shift priorities: technical and social." O'Neil (2016, p. 12) describes the various ways that algorithmic assemblages, in her terms "weapons of math destruction," reflect and enact the biases, prejudicial beliefs, and misunderstandings of the people who developed them. She (2016, p. 10) argues that these math powered applications "define their own reality and use it to justify their results" and further, they tend "to punish the poor and oppressed in our society, while making the rich richer." O'Neil (2016, p. 12) details the negative impacts of algorithmic assemblages in higher education admissions, online, target advertising, criminal justice, the employment application process, surveillance in the workplace, and in the credit and insurance domains. According to Winter (2015, p. 132) "[i]n many cases, algorithmic discrimination unjustly harms individuals or groups who are already socially and economically disadvantaged."

In general, accountability is "the duty to justify a given action to others and be answerable for the results of that action after it has been performed" (Leonelli, 2016, p. 1). To what extent are the developers of algorithms responsible for the future actions of their creations? To what extent do the organizations that employ these developers bear any responsibility for the same? To what extent can the organizations that make use of the algorithmic assemblages (through lease, subscription, or purchase) be held responsible for the decisions made based on algorithmic analysis? These are some of the questions that underlie the concern for algorithmic accountability.

As the consequences of the pervasiveness of algorithmic assemblages in people's work and social lives become clearer, there have been calls to hold accountable the organizations that own, use and/or control these algorithmic assemblages. Such a call is clearly moving away from the utopianism of the dominant technological frame described above. One approach to algorithmic accountability is framed as a technical exercise and involves looking for (Dourish, 2016, p. 6):

> … new ways to make algorithmic processes visible, to render algorithms accountable, and to find within the algorithmic process some opportunity for audit, external review, and examination.

Sandvig, Hamilton, Karahalios, and Langbort (2014, p. 3) similarly argue that "public interest scrutiny of algorithms is required [to] focus on subtle patterns of problematic behavior and that this may not be discernable directly or via a particular instance;" therefore, it is necessary to engage in auditing algorithms. This type of scrutiny would focus on such factors as uncovering the criteria that are coded into prioritization, ranking, sorting, and classification algorithms looking for the extent to which they are "politicized or biased in some consequential way;" the conditions under which algorithms fail; and the presence of bias in the training data (Diakopoulos, 2014, p. 9). It could also involve attempts to reverse engineer algorithms, because "[w]e don't necessarily need to understand the code of the algorithm to start surmising something about how the algorithm works in practice." (Diakopoulos, 2014, p. 14).

Problems quickly arise with this technical exercise approach to algorithmic accountability. For example, there is the opaque nature of algorithms, because if access is gained, as Seaver (2017) notes, algorithms are rarely straightforward to deconstruct. Within code, algorithms are usually woven together with hundreds of other algorithms to create algorithmic systems" (Kitchin, 2017, p. 7). Clearly specialist knowledge is required to unravel the mangle of code to first find the relevant algorithm (Dourish, 2016, p. 6), and second, to isolate and extract the code that is responsible for the biased or otherwise problematic outcome. There is also a question of whether the extracted excerpt of code is actually responsible for the outcome or whether the outcome requires the extracted code acting in concert with other components of the algorithmic assemblage. An additional complexity that may arise if the source of the bias or problem is thought to lie in the training set used to tune the algorithmic system. Assuming that it would be possible to gain access to the training dataset, finding specific evidence of biased data would be a difficult task.

There is also a problem of legal opacity where algorithms have patent or trade secret protection (Barnett et al., 2017, p. 7). Although some have proposed reverse engineering algorithms as a strategy to work around this type of protection, the problem here is that while it might show how the algorithm works in practice, reverse engineering results in little more than speculation about the inner workings of the algorithm (Kitchin, 2017, p. 11), making the attribution of accountability difficult. Algorithmic audits are another proposed method of determining accountability. The question here is whether this type of audit, which focuses on the forensic examination of an algorithm, can generate reliable evidence that bias or discrimination has taken place.

There is also a technical issue posed by the dynamic nature of algorithms. Willson (2016; 148) explains that:

> [a]lgorithms are dynamic processes designed and implemented by humans in conjunction with technical affordances and within broader political, social and cultural environments that are shaped by the continual interactions of strategies, structures and tactics."

What this means is that "algorithms are made and remade in every instance of their use because every click, every query, changes the tool incrementally." (Gillespie, 2014, p. 7). To make matters more challenging, this development takes place over time, as algorithms are coded, reviewed, and tested by teams whose composition changes over time and that do not typically create and maintain clear documentation of the process. Which version of the algorithm is the one for which the organization should be held accountable? If a person is denied a loan because of an algorithmically driven decision and seeks to challenge the outcome, how can the precise version of the algorithm be isolated and examined if it is changing as it is used?

Given the challenges of the technical exercise approach, an alternative approach has been proposed (Musiani, 2013; Neyland, 2019). The sociotechnical approach to algorithmic accountability recognizes that algorithmic assemblages are embedded in complex social, cultural, and organizational contexts where, in addition to the technical aspects of algorithms, the enactment of the algorithm involves the activities of individuals and groups. As Musiani (2013; p.3) explains, the questions on which accountability is based should subtly change:

> By asking questions such as: who are the arbiters of algorithms? Is algorithm design an assertion of authority over more than the algorithm itself? What is the autonomy of algorithms, if any? - it is the accountability and the responsibility of algorithms as socio-technical artifacts that is examined, that of their creators and users, and ultimately, of the balance of power facilitated or caused by algorithms.

At the center of this approach is the insight that there are a number of human influences embedded into algorithms, such as criteria choices, training data, semantics, and interpretation (Diakopoulos, 2014, p. 9; Haarkens, 2018, p. 25; Iliadis, 2018, p. 3). Any investigation must therefore consider algorithms as objects of human creation and take into account intent, including that of any group or institutional processes that may have influenced their design.

To assess algorithmic accountability in this approach would involve "[i]nterviewing designers and coders, or conducting an ethnography of a coding team" because it can provide "a means of uncovering the story behind the production of an algorithm and to interrogate its purpose and assumptions" (Kitchin, 2016, p. 11). Because the team is embedded in a complex organization, it would also be necessary to extent the scope of this research to include relevant stakeholders outside of the technical team as well as a dataset of relevant documentation related to the development of the algorithm.

However, there are a number of difficulties when moving from the technical exercise to the socio-technical approach, and when interviewing individuals and teams of coders to assess sociotechnical algorithmic accountability. In addition to focusing on the coding team, the sociotechnical frame would require extensive investigation of the social and organizational context involving, for example, the analysis of relevant company, advertising and marketing, and legal documents, interviews and observations of key stakeholders in the development, and implementation of the algorithm. This labor and time intensive process cannot happen quickly; Neyland (2019) conducted such an ethnography of a team developing an algorithmically based surveillance system and the work took 3 years. Particularly in the case of (particularly legal) attempts to hold organizations accountable for the outcomes of their algorithmic assemblages, the time required to collet relevant data that can serve as evidence means that the effort has less of a chance of being useful.

RETHINKING ALGORITHMIC ACCOUNTABILITY

A difficulty with the current state of algorithmically-mediated decision making is that according to Leonelli (2016), p. 1), it is:

> difficult to determine who is responsible for what output, and how such responsibilities relate to each other; what ‘participation’ means and which accountabilities it involves, with regard to data ownership, donation and sharing as well as data analysis, re-use and authorship

Barnett et al. (2017, p. 1) argue from a legal standpoint that algorithms are likely to be treated “as artificial persons: a legal entity that is not a human being but for certain purposes is considered by virtue of statute to be a natural person.” The path to accountability then seems clear, meaning that if harm results from the actions of an algorithm, “just sue the humans who deployed the algorithm” (Barnett et al., 2017, p. 7). However, as Barnett goes on to argue, the reality may be more complex - for example, if a person suffers bodily injury as a result of a crash with an autonomous vehicle, who (or what) should be sued? To prove discrimination, it is necessary to demonstrate that the practice or requirement used in the process was chosen because it was likely to have an adverse effect on a person or group (Kleinberg et al., 2019, p. 8). This is an example of disparate treatment. Disparate impact, on the other hand, is discrimination based on an adverse effect against a protected group. In both cases, the person or organization charged with discrimination must justify the choices based on reasons such as “business necessity” or other “neutral” reasons.

One approach to accountability does not focus on the internal workings of the algorithm, thereby avoiding the problem of opening the “black box.” In this view, attention should be turned instead to monitoring the outcomes of algorithmic decision making looking for instance of bias and discrimination, particularly when directed at vulnerable populations (Shah, 2018, p. 3). When considering algorithmic discrimination, in many cases there are two distinct types of algorithms at work; this is typically the case with prediction and selection algorithms. The first is the training algorithm, which analyzes a training dataset looking for patterns and making predictions on the basis of the choices made by the designers and developers. These choices involve the desired outcome to be predicted (hire this person) and the set of variables used to make the prediction (previous salary, education level, relevant experience). When the decision is made that the outputs of this first stage are within the desired parameters, the training algorithm “produces” a screening algorithm, which can then be used on specific cases to make specific predictions. This latter algorithm “cannot do “literally anything” – it is mechanically the result of whatever human decisions were made for the trainer” (Kleinberg et al., 2019, p. 16). Once the decisions have been made about the training data and the desired outcome, the statistical relationships among the variables in the training dataset determine which predictors will be used by the screener algorithm and the weights that they will have in the analysis. Predictions are then made on the basis of these historical data and variables. Therefore, accountability focuses on the social and organizational context of the training algorithm and the training dataset. For this to work, the algorithms and dataset must be “fixed, stored objects that can be inspected” (Kleinberg et al., 2019, p. 17).

Martin (2018, p. 1) advocates for an approach to algorithmic accountability that is tied to the decision; if the algorithm has the main responsibility for making the decision, then the developer should be held accountable. An implication of this approach is that organizations can be held accountable for ethical problems resulting from the operation of their algorithmic assemblages, even if they are working as designed.

In the case of algorithmic assemblages designed to act autonomously, Martin (2018, p. 9) argues that accountability lies with the organization that created and sold or leased the system; in his view, the “inscrutable defense (‘It’s too complicated to explain’) does not absolve a firm from responsibility.” Since the organization and its developers designed the algorithmic assemblage to play the major role in the decision context, they have inserted themselves into the context and have therefore taken on liability for the consequences of their system’s operations. In this scenario, the end users are effectively cut out of the decision making and are simply delivering the algorithm’s output, they cannot be held responsible for the negative consequences of the results.

The ACM (2017) has laid the groundwork for an approach to algorithmically-mediated decision making with its statement on algorithmic transparency and accountability. The seven principles described in the document call on stakeholders to be aware of potential harms that can occur and to provide a means by which affected groups can seek redress for harms that do occur. They suggest that organizations using algorithms should be able to explain to the affected groups how the algorithms work and should be held responsible for the outcomes of algorithmically-mediated decision making. Further, it is the organization's responsibility to provide. Clear documentation of the training, development and testing processes to provide an audit trail. It should be noted that, at this time, these are only guidelines.

CONCLUSION

Despite the technological frame that portrays algorithms and the assemblages in which they are embedded as objective, autonomous, and efficient, it has become increasingly clear that "… algorithms are not neutral but value-laden in that they (1) create moral consequences, (2) reinforce or undercut ethical principles, or (3) enable or diminish stakeholder rights and dignity." (Martin, 2018, p. 4). As algorithmically-mediated decision making becomes more pervasive in the public and private sectors, the potential for discriminatory outcomes also becomes more apparent. Barnett et al. (2017, p. 9) point out that "the lack of redress can have severe consequences affecting individuals as well as groups and whole societies."

Providing digital justice for people and groups who believe they have been harmed as a consequence of algorithmically-mediated decision making is a complex and difficult problem that is attracting the attention of scholars and researchers in many different fields. The main approach to digital justice has been to rely on algorithmic accountability. As a foundation for digital justice, the concept has promise, but as currently conceived, has problems of its own. Decomposed into technical and sociotechnical versions, algorithmic accountability is difficult to achieve. In the case of the former, deconstructing complex and dynamic algorithms, particularly when they are embedded in algorithmic assemblages where they interact with other algorithms, is a difficult and not likely to produce anything like a smoking gun, meaning the lines of code that provide evidence of discrimination (Kleinberg et al., 2019; Seaver, 2017). In the case of the latter, an ethnographic investigation is required that is time-consuming and expensive; it is made more difficult by the fact that algorithmic assemblages are developed and maintained by teams whose composition frequently shifts so again, finding the evidence of discrimination becomes a challenge.

Kleinberg et al. propose imposing a regime of accountability on the developers of algorithmic assemblages whereby they would be required to provide extensive documentation of their design and development processes that would make clear the decisions made when, for example, compiling the data set for training an algorithm. In fact, they would require that the training data set be reserved any be made available upon legal request. Martin (2017) also calls for all approach to algorithmic accountability the shifts the focus from the algorithm to the decision. If an organization provides an algorithm to a third party and has designed it in such a way as to remove the third party from the decision, the organization bears full responsibility for negative outcomes of the decision. Both of these approaches have been reinforced by the ACM, which recently proposes principles for algorithmic accountability and transparency.

Other approaches have also provided glimpses of ways forward. For example, Ananny and Crawford (2018) deconstruct the concept of transparency and in doing so, they provide an extended critique of a central concept in the technical approach to algorithmic accountability. Challenging the assumption that "observation produces insights which create the knowledge required to govern and hold systems accountable," they (2018; 294) argue that transparency is not a state of being able to see clearly, but is, in fact, a means of control that can be harmful, can be used to obscure, and can privilege seeing over understanding. They (2018; 983) also make the point that the central issue is not a matter of holding algorithms accountable - it is more an effort to figure out how to hold algorithmic assemblages accountable which requires "requires not just seeing inside any one component of an assemblage but understanding how it works as a system."

Binns (2018) offers a different way forward, bypassing the technical issues with accountability to move to the step where certain stakeholders are held publicly accountable for their actions. Drawing on political philosophy, he argues for a concept of algorithmic accountability based on the use of public reason, which amounts to a set of laws, beliefs, and regulations acceptable to all of the stakeholders involved in the decision making. The focus here is on the decision makers using algorithmic assemblages during the course of their work and not on the technical details of the algorithms themselves. In what Binns (2018, p. 544) considers the final stage of accountability, four questions must be asked. These revolve around the nature of an acceptable set of justifications about which decision makers and those affected by the decisions can agree, a means for resolving disputed should agreement not be forthcoming, the determinant of precedence when such disagreements occur (meaning, for example that the decision is imposed over the objections of the person affected by the decision), and a means to resolve "legitimate, epistemic and ethical standards to which algorithmic decisions are held." Binns (2018), p. 550) argues that

> public reason could act as a constraint on algorithmic decision-making power by ensuring that decision-makers must be able to account for their system's outputs according to epistemic and normative standards which are acceptable to all reasonable people.

Wong (2019) also focuses on the stage of accountability that occurs after the algorithmic assemblage has generated its output, arguing that the determination of algorithmic fairness is a political rather than a technical concern. This is due, in part, because the (Wong 2019, p. 4) the "idea of "fairness" in algorithmic fairness is in many ways contestable … [which] foregrounds the need to settle the meaning of fairness alongside, if not prior to, the technical tasks" involved in algorithm development. He proposes a different framework, Accountability for Reasonableness, adapted from decision making in health care, as a procedure for deliberating the fairness of any algorithm. Based on the assumption that in a liberal democracy there will always be disagreements between reasonable people, he offers criteria of publicity, relevance, revisions and appeals, and regulation as means by which algorithmic fairness can be ascertained. The rationales and justification for algorithmic decision making should be publicly available (publicity) and should make clear the reasons why the algorithmic assemblage works the way it does and why it is the best way to make decisions about the affected groups (relevance). There should be procedures for challenging these decisions and for dispute resolution that do not privilege either the decision makers or the affected groups ((revisions and appeals) that are monitored and regulated by appropriate entities (regulation).

Rahwan (2018) proposes a sociotechnical means of holding algorithms accountable based on a variation of social contact theory applied to a concept derived from HCI, human-in-the-loop (HITL), which is transformed into society-in-the-loop (SITL). An example of HITL is the case where interactive machine learning systems can speed the process of training when there is regular feedback from human users. Rahwan (2018, p. 7) describes more powerful applications of HITL in algorithmic assemblages; it can

> can also be a powerful tool for regulating the behavior of AI systems … The human can identify misbehavior by an otherwise autonomous system, and take corrective action … The human can be involved in order to provide an accountable entity in case the system misbehaves

To make the move from the human to society, Rahwan (2018, p. 9) proposes SITL, which is HITL and an algorithmic social contract between society and algorithmic assemblages that resolves tradeoffs between competing values on which assemblages can be based such as between privacy and security or different notions of fairness and determines who benefits and who pays the costs of the decisions made. The goal of this approach is to "to build institutions and tools that put the society-in-the-loop of algorithmic systems, and allows us to program, debug, and monitor the algorithmic social contract between humans and governance algorithms."

This paper has engaged in brush clearing to critically examine the concept of algorithmic accountability as a step toward providing a firm grounding for the concept of digital justice by moving away from a focus on the algorithm itself. The next steps involve a review of the range of alternatives that have been proposed.

REFERENCES

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Barnett, J., Koshiyama, A. S., & Treleaven, P. (2017). *Algorithms and the Law*. Legal Futures. Retrieved from https://www.legalfutures.co.uk/blog/algorithms-and-the-law

Barocas, S., Bradley, E., Honavar, V., & Provost, F. (2017). Big data, data science, and civil rights. arXiv preprint arXiv: 1706.03102.

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, *31*(4), 543–556.

Bucher, T. (2012). Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society*, *14*(7), 1164–1180.

Cheney-Lippold, J. (2011). A new algorithmic identity: Soft biopolitics and the modulation of control. *Theory, Culture & Society*, *28*(6), 164–181.

Couldry, N., Gray, M. L., & Gillespie, T. (2013). Culture digitally: Digital in/justice. *Journal of Broadcasting & Electronic Media.*, *57*(4), 608–617.

Diakopoulos, N. (2014). *Algorithmic accountability reporting: On the investigation of black boxes*. Tow Center for Digital Journalism, Columbia University.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62.

Dourish, P. (2016). *Algorithms and their others: Algorithmic culture in context*. Big Data and Society.

Fink, K. (2017). Opening the government's black boxes: Freedom of information and algorithmic accountability. *Information, Communication & Society*, 1–19.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot, *Media technologies* (pp. 167–194). Cambridge, MA: MIT Press.

Haarkens, A. (2018). The ghost in the legal machine: Algorithmic governmentality, economy, and the practice of law. *Journal of Information, Communication and Ethics in Society*, *16*(1), 16–31.

Iliadis, A. (2018). Algorithms, ontology, and social progress. *Global Media and Communication*, 1–12.

Jentile, C., & Lawrence, M. (2016). *How government use of big data can harm communities*. Ford Foundation.

https://www.fordfoundation.org/ideas/equals-change-blog/posts/how-government-use-of-big-data-can-harm-communities/

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, 1–16.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). *Discrimination in the age of algorithms* (No. w25548). National Bureau of Economic Research

Klett, J. (2016). Baffled by an algorithm: Mediation and the auditory relations of 'immersive audio'. In R. Seyfert & J. Roberge, *Algorithmic cultures: Essays on meaning, performance, and new technologies* (pp. 111–127). London, England and New York, NY: Routledge.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1–16.

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A*, *374*(2083), 20160122.

Lustig, C., & Nardi, B. (2015, January). Algorithmic authority: The case of Bitcoin. In *Proceedings of the 48th Hawaii International Conference on the System Sciences* (pp. 743–752). IEEE.

Martin, K. (2018). Ethical implications and accountability of algorithms. *Journal of Business Ethics*.

McCann, D., Hall, M., & Warin, R. (2018). Controlled by calculations? Power and accountability in the digital economy. *New Economics Foundation.*

McCarthy, M. T. (2016). The big data divide and its consequences. *Sociology Compass*, *10*(12), 1131–1140.

McQuillan, D. (2017). Data science as machinic neoplatonism. *Philosophy & Technology*, 1–20.

Musiani, F. (2013). Governance by algorithms. *Internet Policy Review*, *2*(3), 1–8.

Neyland, D. (2019). *The everyday life of an algorithm*. Springer International Publishing.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Prins, C. (2018). Digital justice. *Computer Law & Security Review*, *34*(4), 920–923.

Rainie L., & Anderson J. (2017). *Code-dependent: Pros and cons of the algorithm age*. Pew Research Center. Retrieved from http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/

Sabelli, C., & Tallacchini, M. (2018). From privacy to algorithms' fairness. In M. Hansen, E. Kosta, I. Nai-Fovino, & S. Fischer-Hübner, *Privacy and identity management. The smart revolution. privacy and identity 2017 IFIP advances in information and communication technology* (Vol. *526*, pp. 86–110). Cham, IL: Springer.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Presented at. In *Data and discrimination: Converting critical concerns into productive inquiry* (pp. 1–23).

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 1–12.

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A*, *376*(2128), 20170362.

Striphas, T. (2015). Algorithmic culture. *European Journal of Cultural Studies*, *18*(4-5), 395–412.

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), 1–14.

Willson, M. (2016). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150.