# ALGORITHMIC ACCOUNTABILITY *IN THE MAKING*

By Deborah G. Johnson*

Abstract: *Algorithms are now routinely used in decision-making; they are potent components in decisions that affect the lives of individuals and the activities of public and private institutions. Although use of algorithms has many benefits, a number of problems have been identified with their use in certain domains, most notably in domains where safety and fairness are important. Awareness of these problems has generated public discourse calling for algorithmic accountability. However, the current discourse focuses largely on algorithms and their opacity. I argue that this reflects a narrow and inadequate understanding of accountability. I sketch an account of accountability that takes accountability to be a social practice constituted by actors, forums, shared beliefs and norms, performativity, and sanctions, and aimed at putting constraints on the exercise of power. On this account, algorithmic accountability is not yet constituted; it is in the making. The account brings to light a set of questions that must be addressed to establish it.*

KEY WORDS: accountability, algorithms, algorithmic accountability, artificial intelligence, algorithmic decision-making, social norms

## I. Introduction

Algorithms are now used in an ever-expanding number of domains to make decisions that powerfully affect the lives of individuals and the activities of private and public institutions. To name a few of these domains, algorithms are used to decide who gets loans and mortgages, what sentences are given to convicted criminals and who is released from prison on probation, what job advertisements are shown to which group of potential candidates, how colleges are ranked, how a car reacts to an object in its path, what music or video recommendations are provided to users, what treatments are recommended for patients, and on and on. In defense of these uses of algorithms, many point out that human decision-making capabilities are limited in certain domains and often prone to error and bias, so in certain contexts algorithms can do better than humans. Indeed, there are contexts in which AI algorithms can make determinations that could not have been made by humans, for example, in detecting very small tumors in

* Department of Engineering and Society, University of Virginia, dgj7p@virginia.edu. Competing Interests: The author declares none. In writing this essay, I benefited enormously from comments provided by Missy Cummings, Jack Davidson, Nick Diakopoulos, Geeta Patel, and Mario Verdicchio and from a thorough vetting of the essay by reviewers and the other contributors to this volume.

doi:10.1017/S0265052522000073

medical imagery[1], fraud detection in online transactions,[2] and optimizing supply chains and control inventory in manufacturing.[3]

Nevertheless, a wide variety of concerns have been expressed about algorithmic decision-making. Perhaps most worrisome are concerns about safety and unfair bias. The safety issue has come to the fore prominently in discourse about autonomous vehicles. The ultimate success of autonomous cars depends on the development of AI systems that will assist in various functions of the car, including the systems that detect objects in a vehicle's path. There have already been several accidents with fatalities that are traced back to failures in AI. This has drawn attention to the trustworthiness and reliability of AI decision-making. The challenge here is not just to ascertain when the cars will be safe enough but to determine how safety can be measured.[4]

In addition to safety, a good deal of public attention has been given to bias in algorithms.[5] Algorithms have been found, in multiple contexts, to produce racially- and gender biased decisions. In criminal justice, for example, racial bias has been detected in the algorithm-based systems that identify where crimes are likely to occur and which individuals are likely to commit them.[6] Racial bias has also been detected in algorithm-based systems that provide sentencing advice to U.S. correctional offender boards.[7] In hiring and recruitment, gender bias has been detected in algorithmic systems that are used to advertise job openings.[8] In loan and mortgage lending, the discriminatory outcomes of algorithm-based systems have been described as algorithmic redlining.[9] Bias can be introduced in multiple ways; Danks and London distinguish training data bias, algorithmic focus bias; algorithmic processing bias; transfer context bias; and interpretation bias.[10] Cummings explains that the design of algorithms involves subjective (and therefore potentially biased) decision points, including in the selection of

[1] Greg Freiherr, "AI Algorithm Detects Breast Cancer in MR Images," https://www.itnonline.com/article/ai-algorithm-detects-breast-cancer-mr-images-0

[2] Niccolo Mejia, "AI-Based Fraud Detection in Banking—Current Applications and Trends," https://emerj.com/ai-sector-overviews/artificial-intelligence-fraud-banking/

[3] Philip Kushmaro, "5 Ways Industrial AI is Revolutionizing Manufacturing," https://www.cio.com/article/3309058/5-ways-industrial-ai-is-revolutionizing-manufacturing.html

[4] Mary L. Cummings and Jason Ryan, "Point of View: Who Is in Charge? The Promises and Pitfalls of Driverless Cars," *TR News* 292 (2014).

[5] See, for example: Megan Garcia, "Racist in the Machine: The Disturbing Implications of Algorithmic Bias," *World Policy Journal* 33, no. 4 (2016): 111–17.

[6] Elizabeth E. Joh, "Policing by Numbers: Big Data and the Fourth Amendment," *Washington Law Review* 89, no. 1 (2014): 35–68.

[7] Ansgar Koene, "Algorithmic Bias: Addressing Growing Concerns [Leading Edge]," *IEEE Technology and Society Magazine* 36, no. 2 (2017): 31–32.

[8] Anja Lambrecht and Catherine Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *Management Science* 65, no. 7 (2019): 2966–81.

[9] Michelle Chen, "Redlined by Algorithm," https://www.dissentmagazine.org/online_articles/redlined-by-algorithm

[10] David Danks and Alex John London, "Algorithmic Bias in Autonomous Systems," *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (2017): 4691–97.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

data sets, the process of modifying data sets to make them compatible, deciding when the model is accurate, choosing which features of the output are significant, and in assigning weights to features.[11]

As a result of these concerns, robust public discourse has called for *algorithmic accountability.* The current discourse focuses largely on algorithms and their opacity, and I will argue here that this reflects a narrow and inadequate understanding of accountability. The discourse would benefit from a broader discussion of the nature of accountability and how it works. In what follows, after explaining how and why attention is being given to algorithms and their opacity, I provide a broad account of accountability that takes it to be a social practice involving actors, forums, shared beliefs and norms, performativity, and sanctions. This broader account is then extended to algorithmic accountability. The thrust of the argument is that while attention to the opacity of algorithms is not a bad thing, more attention should be focused on the actors, forums, and norms needed to constitute algorithmic accountability.

One clarification of language will be helpful for those not familiar with algorithms. In the discourse on algorithmic accountability, the terms "algorithmic accountability" and "AI accountability" are used inconsistently. AI algorithms can be understood to be a particular kind of algorithm. An algorithm is a sequence of computational steps that transform input into output. While programmers typically know the steps that an algorithm directs a computer to go through to produce output, the programmers of AI algorithms may not know exactly how their algorithm achieves results.[12] AI algorithms use AI techniques and many AI algorithms in use today are based on machine learning (ML). These algorithms take large data sets (training data) as input and, based on a model, use the data to figure out *how* to achieve the desired form of output. Once trained, the AI algorithms produce output that is used. The classic example is teaching a program to identify cats by giving it pictures of cats and things that aren't cats (the training data) and telling it which are cats. The AI learns how to identify cats and, when given new images, can identify which are cats and which not. However, the programmers do not know exactly how the AI algorithm achieves its results.

Of course, the AI systems being developed today perform much more complicated and important tasks than identifying cats. As already explained, current AI systems are used for a wide range of activities from scanning images to identify tumors, to predicting which prisoners are likely to commit crimes again, to identifying who is more and less likely to pay back loans, be successful in school, be a good dating match, and so on.

[11] M. L. Cummings, "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings," *AI Magazine* (March 22, 2021).

[12] Mike Ananny and Kate Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media and Society* 20, no. 3 (2018): 973–89.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

Although AI algorithmic decision-making is an especially challenging case, the account of accountability presented here applies broadly to AI and non-AI algorithmic decision-making.

## II. Opacity as the Challenge of Algorithmic Accountability

The early literature on AI and accountability emphasized a "responsibility gap" in AI decision-making due to the fact that the humans who created the algorithms could not understand *how* the algorithms achieved their results (outputs) and, therefore, could not be held accountable.[13] Recently, however, the discourse seems to have given way to more pragmatic attempts to develop accountability.[14] In part at least, this seems to arise from recognition that AI might be rejected unless an appropriate form of accountability is instituted. For example, Dolshi-Velez notes that the "question of how to hold AI systems accountable is important and subtle: poor choices may result in regulation that not only fails to truly improve accountability but also stifles the many beneficial applications of AI systems."[15]

Currently, the discourse has largely identified the *opacity* of algorithms as the crux of the problem.[16] The logic and steps that algorithms go through to produce output is opaque—to those who use the algorithms in decision making, to those who are affected by the decisions, and in the case of AI algorithms even, as just explained, to those who have designed the algorithms. This opacity is partly due to complexity and scale (that is, the size of data sets and the number of data attributes processed by an algorithm) but in the case of AI algorithms, it is also the fact that the algorithm *learns* as it operates, and the programmer does not understand exactly how or what the AI algorithm learns. Concerns about safety and bias are intertwined with opacity in the, perhaps obvious, sense that if you can't tell how something works, it is difficult, if not impossible, to know that it will be safe in all circumstances or that it won't learn to rely on some element in data that skews racially or by gender. Arguably, there are ways to test for safety and bias and ways to protect against unsafe and

[13] See Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6, no. 3 (2004): 175–83; Rob Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77; and Deborah Johnson, "Technology with No Human Responsibility?" *Journal of Business Ethics* 127, no. 4 (2015): 707–15.

[14] See, for example, Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making," *Communications of the ACM* 59, no. 2 (2016): 56–62.

[15] Finale Doshi-Velez and Mason Kortz, "Accountability of AI Under the Law: The Role of Explanation," Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet and Society working paper (2017), 2, http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

[16] See, for example, J. Burrell, "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms," *Big Data and Society* 1 (2016): 1–12; and Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein, "Discrimination in the Age of Algorithms," *Journal of Legal Analysis* 10 (2018): 113–74.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

biased decisions that don’t require full understanding, but these strategies are not always foolproof, and in the current state of algorithm development are not well understood.

With opacity understood to be the problem, not surprisingly, transparency is predominantly seen as the solution.[17] Indeed, in the discourse on algorithmic accountability, accountability is frequently equated with transparency.[18] Transparency is a common trope for accountability. The core idea of transparency is that “sunlight disinfects.” Actors are less likely to engage in bad behavior when they are compelled to act in full sight. In this respect, accountability is preventative as well as post ante. The inference for algorithms is that dangerous or biased algorithms are less likely to be developed and used if those who design and use them are compelled to show how they work.

Although transparency figures largely in the discourse on algorithmic accountability, the challenges of achieving transparency are also identified.[19] Transparency is relational in the sense that it is a matter of revealing or providing information to an individual or group, and the information has to be in a form that the individual or group can understand and use. In the case of algorithms, complexity and scale make it difficult for non-experts to understand, and in the case of AI algorithms, as already explained, even experts admit that they can’t fully understand how algorithms work at a granular level.

In the discourse on algorithmic accountability, transparency is often equated with explainability. This move has been reinforced (if not generated) by the European Union General Data Protection Regulation (EU GDPR), which specifies that individuals have a right to an explanation when they are turned down for benefits in processes that involve automated decision making.[20] The GDPR specifies that individuals have a right to “meaningful information about the logic involved” in automated decision-making.[21] Although the United States has nothing as concrete as the

[17] Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu, “Accountable Algorithms,” *University of Pennsylvania Law Review* 165, no. 3 (2017): 633–706; L. M. Cysneiros, M. Raffi and J. C. Sampaio do Prado Leite, “Software Transparency as a Key Requirement for Self-Driving Cars,” *2018 IEEE 26th International Requirements Engineering Conference (RE)* (2018): 382–87; Diakopoulos, “Accountability in Algorithmic Decision Making.”

[18] See, for example, Paul B. De Laat, “Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?" *Philosophy and Technology* 31, no. 4 (2018): 525–41; and Nicholas Diakopoulos, “Algorithmic Accountability,” *Digital Journalism* 3, no. 3 (2015): 398–415.

[19] Ananny and Crawford, “Seeing Without Knowing”; Kroll et al., “Accountable Algorithms,” and Nicholas Diakopoulos, “Accountability in Algorithmic Decision-Making,” *Queue* 13, no. 9 (2015): 126-49.

[20] EU GDPR articles 13–15, https://gdpr.eu/tag/gdpr/

[21] For discussion of these GDPR articles, see A. D. Selbst and J. Powles, “Meaningful Information and the Right to Explanation,” *International Data Privacy Law* 7, no. 4 (2017): 233–42; and Bryce Goodman, and Seth Flaxman, “European Union Regulations on Algorithmic Decision-Making and a ‘Right to Explanation,’” *AI Magazine* 38, no. 3 (2017): 50–57.

GDPR, explainability is a prominent theme in addressing algorithmic accountability. For example, the U.S. Association for Computing Machinery (ACM) policy statement "Principles for Algorithmic Transparency and Accountability" uses explainability.[22] The third of the five principles under the label "Accountability" specifies that "Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results." The fourth principle then specifies that "Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made."[23]

Technologists have taken up the challenge of explainability by pursuing technological ways to produce explanations. A new line of research has been spawned that focuses on techniques to design algorithms that generate explanations of their operations.[24] So far it seems unclear what criteria will be used to determine what counts as an adequate explanation and for whom.

Transparency is by no means a simple concept.[25] In the case of algorithms, it would require those who produce algorithms to explain them to a particular audience or audiences. However, corporate secrecy laws are likely to limit what algorithm designers can be required to reveal.[26] Moreover, transparency always involves decisions about what to divulge, in what form, where, and what to leave out,[27] so there is no guarantee that information revealed in the name of transparency will provide the kind of information needed for accountability for bad decisions.

Although attention to algorithms and their opacity is not entirely misguided, to think that algorithmic accountability can be achieved merely or even primarily by making algorithms transparent is. To understand why this is so, we must delve more deeply into the nature of accountability and how it operates to achieve its function.

[22] The ACM is the largest computing society in the world.

[23] Statement on Algorithmic Transparency and Accountability, January 12, 2017, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[24] Doshi-Velez and Kortz, "Accountability of AI Under the Law"; Finale Doshi-Velez and Been Kim, "Towards a Rigorous Science of Interpretable Machine Learning," https://arxiv.org/abs/1702.08608; Wojciech Samek and Klaus-Robert Müller, "Towards Explainable Artificial Intelligence," in W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K. R. Müller, eds., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, vol. 11700 (Cham: Springer, 2019), 5–22.

[25] Nicolas Diakopoulos, "Transparency," in Markus Dubber, Frank Pasquale, and Sunit Das, eds., *Oxford Handbook of Ethics and AI* (New York: Oxford University Press, 2020).

[26] AI Now 2018 Report, https://ainowinstitute.org/AI_Now_2018_Report.pdf

[27] Archon Fung, Mary Graham, and David Weil, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge: Cambridge University Press, 2007).

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

## III. A Broader Account of Accountability

Mark Bovens provides an account of accountability that captures much of what is essential to the concept.[28] He describes accountability as "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences."[29] This account is a meticulous specification of a consensus view. That is, the elements in Bovens's account appear and reappear in discussions of accountability. For example, Grant and Keohane explain that accountability "implies that some actors have the right to hold other actors to a set of standards, to judge whether they have fulfilled their responsibility in light of these standards, and to impose sanctions if they determine that these responsibilities have not been met."[30] Schedler reduces the elements of Bovens's account to just two, claiming that (political) accountability has two connotations: "answerability, the obligation of public officials to inform about and to explain what they are doing; and enforcement, the capacity of accounting agencies to impose sanctions on powerholders who have violated their public duties."[31]

Although Boven's focus (and much of the literature on accountability) is on accountability in political contexts (for instance, the accountability of political leaders, public administrative units, and nation-states), his account has broader application and is consistent with accountability in a wide variety of contexts. For example, in criminal justice, individuals (actors) are understood to have an obligation to appear in court and explain or justify their behavior when it appears to break the law. They present themselves before a judge and/or jury that represents the public (the forum). If the explanation is not satisfactory, the judge or jury may find the individual guilty and render appropriate consequences. The same essential elements are present in informal one-on-one (as opposed to one to many) moral practices, for example, when a person discovers evidence of a friend's betrayal (such as revelation of an intimate secret to a third party) and demands an explanation/justification for the seeming betrayal. Friends generally share the belief that they are obligated to explain this kind of behavior. The slighted person may accept the explanation as adequate or find it lacking and, if the latter, may render consequences to the friend (for

[28] Mark Bovens, "Analysing and Assessing Accountability: A Conceptual Framework," *European Law Journal* 13, no. 4 (2007): 447–68. The account also appears in Mark Bovens, T. Schillemans, and R. E. Goodin, "Public Accountability," *The Oxford Handbook of Public Accountability* 1 (2014): 1-20.

[29] Bovens, "Analysing and Assessing Accountability," 450.

[30] Ruth W. Grant and Robert O. Keohane, "Accountability and Abuses of Power in World Politics," *American Political Science Review* 99, no. 1 (2005): 29.

[31] Andreas Schedler, "Conceptualizing accountability," in *The Self-Restraining State: Power and Accountability in New Democracies*, ed. Andreas Schedler, Larry Diamond, and Marc F. Plattner (Boulder, CO: Lynne Rienner Publishers, 1999), 14.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

instance, breaking off the relationship). The same sort of practice is evident when a company is accused of social or environmental wrongdoing (for instance, unsustainable environmental behavior or exploitative labor practices). Here customers or citizen activist groups (forums) may call the company (the actor) to account by demanding an explanation/justification of its behavior, and if the explanation is not adequate or a change in behavior is not promised, the groups may render consequences such as boycotts or negative publicity.

Given the breadth of application of Bovens's account, it seems plausible to use it in thinking about algorithmic accountability. Before doing this, however, the elements of the account need to be carefully examined and this will lead to some modifications. First, the overarching frame of Bovens's account is that accountability is a relationship between an actor and a forum. So, what kinds of entities can constitute this relationship? The *actor* can be an individual or a collective entity such as a corporation, government agency, or profession. A *forum* can, in principle, be any group of individuals who believe an actor has an obligation to account to them. More on this later. Actors may be simultaneously accountable to multiple forums. A company, for example, may be accountable (for different types of behavior) to a board of directors, regulatory agencies, employees, and customers.

A central aspect of the relationship between actor and forum is that both parties share the belief that the actor has an obligation to explain particular types of behavior. This sharing of belief was evident, for example, in the recent Volkswagen emissions fraud scandal when there was consensus among a wide range of forums (for example, customers, regulatory agencies, the automotive industry, and media) as well as the company itself, that the company had an obligation to explain its apparently fraudulent behavior.[32] However, there are situations in which actor and forums disagree. This situation typically occurs when an actor does not believe that an explanation is owed while a forum does. Consider the case of a president of a country who thinks he or she has no obligation to respond to public calls for an explanation/justification of some behavior. On Bovens's account it would seem that we would have to say that there is no accountability because the actor and forum do not share the belief that an explanation is owed. This, however, doesn't seem right. That is, it seems wrong to say that when an actor doesn't believe he or she owes an explanation, the actor is, therefore, not accountable. Among other things, this would eliminate the possibility that an actor could defy accountability. There is also the more complicated situation in which forums disagree about the actor's obligation. Suppose a local citizens' group demands information about the chemicals being released into the local water supply by a manufacturing company, and the company refuses to provide such information on grounds that it has

[32] Russell Hotten, "Volkswagen: The Scandal Explained," *BBC News*, December 10, 2015, https://www.bbc.com/news/business-34324772

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

no legal obligation to do so. Insofar as the law represents a broad public forum, we have here a case in which the actor and one forum agree that an explanation is not owed and another forum disagrees and insists that an explanation is owed. Again, in this situation it seems wrong to say simply that the actor is not accountable.

These problematic cases arise because Bovens's account is descriptive; it is a description of established accountability practices. His account is not focused on how accountability comes to be or how actors and forums come to share beliefs about the obligation to explain. This is important here both because it helps to explain the cases of disagreement just mentioned and because algorithmic accountability is not yet well established.

In well-established accountability practices, as in the Volkswagen emissions fraud case just mentioned, actors and forums share the belief that an explanation is owed. The belief is embedded in social norms and expectations and, in this case, has the backing of law. By contrast, when accountability practices are not fully developed, forums may call for an actor to explain certain behaviors, but the validity of the demand may "fall on deaf ears." The belief that an explanation is owed is not well accepted in such cases. Importantly, when such forums call for explanations, this may initiate or contribute to the development of new accountability practices. Public demands for an explanation can help to create beliefs that an explanation is owed. This will be discussed further when the performative aspect of accountability is discussed. For now, it is important to note that when actors and forums disagree, it does not mean necessarily that there is no accountability, rather it may be a sign of accountability practices *in the making*.

Notice that I have been referring to accountability as if it is a practice. Although Bovens frames accountability as fundamentally a relationship between an actor and a forum, it seems better to frame it as a practice, one that has at its heart the relationship between an actor and a forum. The practice consists of the activities of: calling for an explanation from certain actors for their behavior; engaging in a process of listening to, evaluating, and interrogating the explanation provided; and imposing sanctions when appropriate. In other words, accountability is a practice that exhibits the elements identified by Bovens. Importantly, such practices don't come out of nowhere. They can develop over time through the informal demands of forums, changes in the attitudes and beliefs of actors and forums, and through formal actions instituting, for example, regulation, legislation, and policies.

Another caveat in regard to Bovens's account should be noted. Bovens specifies that actor and forum share *beliefs* but the sharing here need not be consciously- or thoughtfully held beliefs. The idea that an actor has an obligation can be implicit in social norms and expectations. Social norms are patterns of belief and behavior that people adhere to, often without much thought, so they are not beliefs in the sense of strongly held convictions. Indeed, individuals are often not aware of social norms they adhere to

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

until someone violates them. Moreover, in accountability practices it is not just the belief about the obligation to explain that may be implicit. Actors and forums may operate on assumptions about a range of matters, including when an explanation is owed (for what types of behavior), when an explanation is exculpatory, and what are appropriate consequences.

Thinking in terms of social norms allows us to consider the formal and informal means by which shared beliefs are established. Arguably, the most effective means are legal and regulatory specification. For example, the GDPR, mentioned earlier, established a forum's *legal* right to an explanation. Informal norms are implicitly held. Remember the case, mentioned earlier, of a friend's seeming betrayal. In that type of situation, individuals have the expectation of an explanation not because there is a law but because the norms with regard to friendship are part of their understanding of the world, their assumptions about how the social world works. Though such norms are informal, they can be developed intentionally. This is evident in discussions about changing the culture in this or that context. Changing culture means changing the prevailing norms and beliefs about what to expect in a particular environment.

Importantly, when it comes to accountability involving technology, *technological* as well as social norms play a role, and there are typically layers of accountability that each involves norms. For example, in the case of a bridge collapse, the public (the forum) expects (believes that it has a right to) an explanation from relevant authorities as to why the bridge collapsed and what will be done about it. In addition to the private company or government agency that initiated or paid for the construction of the bridge, a design firm, contractors, manufacturers, and inspectors will come into focus. An investigation would be made to determine the cause of the collapse, and this investigation would identify the relevant actors and rely on formal and informal norms applicable to their behavior. Suppose the investigation concluded that a particular beam failed. This would then lead to questions such as the following: Why did the beam fail? Was it flawed? Did those who designed the bridge specify the wrong dimensions for a beam placed there? Was the beam properly installed? Was the amount of pressure put on the beam more than anticipated by the designers? Each one of these questions points to a different norm associated with a different activity and accountable entity—the standards for beam manufacturing, the expectations for those who install beams, how design engineers are to calculate the dimensions of beams used in certain places, who controls the use of the bridge and how they are supposed to prevent too much weight from being put on it. And so on.

The bridge example illustrates both the multiple actors and multiple norms that come into play in accountability practices for outcomes that have a technological component. In the next section, parallel complexities will be seen in the case of algorithmic accountability.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

Another aspect of accountability that comes into clearer view when accountability is framed as a social practice—one that can be in various stages of development—is its *performative dimension.* Bovens doesn't mention this, though it is not inconsistent with his account. The norms implicit in an accountability practice can be established and reinforced and transmitted through performance. In the action of calling an actor to account, receiving their explanation, and rendering consequences, accountability is performed. Moreover, engaging in these activities even when there is no broad consensus about their appropriateness can help to establish new norms and shared beliefs about accountability.

Performativity helps to explain the cases mentioned earlier when actors do not share the belief that they owe an explanation or when forums disagree about the accountability of an actor. The call by a forum for an explanation pressures the actor and promotes the belief that an explanation is owed. Consider the case of corporate social responsibility mentioned earlier. When companies respond to calls for corporate social responsibility by making reports on their activities, even if they fail to address the particular activities for which they are being called to account (for example, if they discuss their charitable activities instead of their degradation of the environment), their reporting performs accountability. The performance reinforces the idea that the forum is entitled to an explanation. When Richard Nixon famously said in response to the Watergate break-in, "I am responsible but not to blame," he performed the ritual of accountability while deflecting the significance of what he had done. So, established accountability practices are performed, and performances of accountability can help to create accountability practices where they are not fully established.

The final element of Bovens's account is sanctions, the consequences a forum may render. This aspect of accountability cannot be fully understood without addressing the function of accountability. Having recognized that accountability is a social practice, we can ask why such practices exist? What purpose do they serve? Accountability is generally understood to be directed at constraining power. It puts constraints on the ability of actors to exercise their power. Schedler explains how this works: "Rather than denoting one specific technique of domesticating power, it [accountability] embraces three different ways of preventing and redressing the abuse of political power. It implies subjecting power to the threat of sanctions; obliging it to be exercised in transparent ways; and forcing it to justify its acts."[33]

Notice that Schedler sees the threat of sanctions as both "preventing and redressing" the abuse of power. Accountability is both forward- and backward-looking. Accountability practices can deter actors from abusing their power as well as respond to those who have not been deterred. These two functions work together in the sense that when the backward-looking practice is performed (for example, an actor is held to account for their bad

[33] Schedler, "Conceptualizing Accountability," 14.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

behavior), this reinforces the threat of sanctions for that behavior and thereby may deter (prevent) others from engaging in that kind of behavior in the future.

Two final caveats should be noted before turning attention fully to algorithmic accountability. First, it may be helpful to keep in mind a distinction between formal and informal practices of accountability. Accountability is formal when the obligation to explain is explicitly and publicly stated in law, regulations, and policies. Statements of this kind intentionally produce shared beliefs and norms. Informal accountability can be as effective, but relies on implicit social norms that, although widely accepted, may not be explicitly discussed. The mere threat of formal accountability practices can pressure actors to constrain their behavior, for example, by adopting self-regulatory practices. Industries often adopt standards and disseminate best practices to fend off more formal forms of accountability. Something like this is underway with regard to social media platforms as they make changes in how they filter misinformation so as to fend off government regulation. The same kind of threat seems to be at work in the discourse on algorithmic accountability. This is evident in the statement quoted earlier by Dolshi-Velez suggesting that some form of accountability practice is needed to fend off regulation that might diminish beneficial applications of AI systems.[34] Dolshi-Velez can be read here as arguing for the development of informal forms of accountability to fend off formal accountability.

Second, since accountability is a practice aimed at constraining power, there may be other strategies to achieve the same function. Zimmermann, Di Rosa, and Kim argue that concerns about algorithmic decision making might be better addressed ex ante (before such decision making power is deployed).[35] They seem to see accountability as strictly a post ante method, only coming into play after undesirable behavior has occurred. Following Frank Pasquale,[36] they argue for an approach that questions whether algorithms should be used at all in certain domains, and they call for public involvement in the design of algorithms and substance of algorithms. This ex ante and democratic approach is a good and important idea, though it is not exactly an alternative to accountability. Rather, ex ante public decisions about the design and use of algorithms would provide norms for accountability practices. In other words, were public decisions made to prohibit the use of algorithms or to restrict the design and use of algorithms in certain domains, this would create norms that would be used to hold actors accountable.

[34] Doshi-Velez and Kortz, "Accountability of AI Under the Law."

[35] A. E. Zimmerman, E. Di Rosa, and H. Kim, "Technology Can't Fix Algorithmic Injustice," *Boston Review: A Political and Literary Forum. Boston Review* https://www.bostonreview.net/science-nature-politics/annette-zimmermann-elena-dirosa-hochan-kim-technology-cant-fix-algorithmic. 2020

[36] Frank Pasquale, "The Second Wave of Algorithmic Accountability," https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability/

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

So, accountability is a social practice involving actors, forums, shared beliefs and norms, performativity, and sanctions, and the practice is aimed at putting constraints on the exercise of power. As such, accountability practices can be in various stages of development depending on how well formed the forums and how well inculcated the norms—be they formal or informal. We can now extend this understanding of accountability to algorithmic accountability.

## IV. Implications for Algorithmic Accountability

When the account is extended to algorithmic accountability, a set of questions comes into view that are rarely addressed in the current public discourse.

1) Who are the *actors* and *forums* in algorithmic accountability?[37]
2) What are the *shared beliefs* and *norms* that constitute algorithmic accountability?
3) Where and how is algorithmic accountability performed?
4) What sanctions are threatened and/or imposed when actors are called to account for algorithmic decision-making?

However, these questions ask for the elements of an existing practice of accountability, and algorithmic accountability is not a well established practice or set of practices. Indeed, the fact that answers to these descriptive questions are not obvious supports the claim that algorithmic accountability is only nascent. There are calls for it, the need has been recognized, the discourse about its shape is robust, but the parameters of the practice are inchoate. Algorithmic accountability is *in the making.*

In the interest of establishing algorithmic accountability, the four questions can be reformulated into normative questions. In this form, they give direction to the endeavor to develop algorithmic accountability practices:

1) Who should be identified and specified as the actors to be held accountable in algorithmic decision-making? What groups are appropriate forums for algorithmic accountability?
2) What shared beliefs and norms should be developed and promulgated?
3) Where and how should algorithmic accountability be performed?
4) What sanctions should be imposed on the actors in algorithmic decision-making?

[37] For a discussion of the importance of critical forums, see Jakko Kemper and Daan Kolkman, "Transparent to Whom? No Algorithmic Accountability without a Critical Audience," *Information, Communication and Society* 22, no. 14 (2019): 2081–96.

Providing answers to these questions is enormously challenging because algorithmic decision-making takes place in so many domains and generally involves *many hands* operating in complex organizational environments. When fully established, algorithmic accountability will likely look more like accountability for bridges than, say, the accountability of a political leader. That is, it is likely to have multiple layers and to be reliant on norms that have been formally and informally inculcated and have come to be widely accepted. Although the four normative questions cannot be fully answered here, some starting places can be identified.

Who should be specified as the actors to be held accountable in algorithmic decision-making? For a start, it is important to note that although algorithms are essential components in algorithmic decision-making, they cannot be considered actors in accountability practices. This may seem too obvious to mention, however, the current discourse on algorithmic accountability emphasizes algorithms and their transparency as if they are the key to accountability. In contrast, the account of accountability just provided does not mention them. Making algorithms transparent may be helpful in holding human actors to account, but attention to algorithms alone is a false path to achieving accountability.[38]

In general, when it comes to outcomes instrumented with technological artifacts, the artifacts are efficacious in producing the outcome, that is, the outcomes could not have been achieved without the efficacy of the artifacts. However, the efficacy of the artifacts only works in combination with human behavior. Algorithms only operate when humans create and use them. For example, in the case of algorithm-based parole decisions, it is the behavior of those who develop the algorithms, those who organize and create databases of information on criminals, those who are trained to use software to generate probabilities for particular individuals, members of parole boards, and others that combine with the efficacy of the algorithm to produce a parole decision. The algorithm has no meaning or efficacy without the behavior of human actors.

Moreover, algorithms can't be actors in accountability practices because they are not capable of having beliefs and obligations, not capable of giving explanations of their behavior, responding to interrogation, and facing consequences.[39] The central actors in algorithmic accountability are algorithm designers and users because they are capable of having beliefs and obligations, adhering to norms, and responding to the threat of sanctions. Holding designers and users of algorithms accountable aligns with the

[38] See Julia Powels and Helen Nissenbaum, "The Seductive Diversion of 'Solving' Bias in Artificial Intelligence," OneZero December 7, 2018, https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53

[39] To be sure, some may argue that designing an algorithm to provide an explanation of its behavior is a way of assigning it the obligation to explain. However, this is true only in a metaphorical sense. Attributing obligations to algorithms is a playful way of using language. The parallel between algorithms and people works as playful against the backdrop of the standard social meaning of artifacts as things that cannot have beliefs or obligations.

function of accountability to constrain power. Designers have the power to make algorithms and users have the power to use algorithms, and their power can be constrained by holding them to account.

Of course, specifying that the central actors in algorithmic accountability must be the designers and users of algorithms only gets us so far since these actors are often organizations with complex arrangements of individuals in varying roles and with varying degrees of power and authority. Designers may be, for example, small academic research groups or large software companies, and users may be individuals or multiplex organizations such as parole boards, insurance companies, and social media platforms. Nevertheless, these actors—large and small—are capable of being held to account; they are capable of instituting and adhering to policies and practices that constrain their power.

The complex organizational context of algorithmic design and use means that algorithmic accountability is not likely to be a single practice but rather a set of practices targeted to the many layers of activity within an organization. For example, there will have to be practices that hold companies, government agencies, and industries to account, as well as practices that hold individuals and groups within those organizations to account. Moreover, practices are likely to vary from domain to domain; that is, accountability is likely to be different when it comes to, for example, algorithmic decision-making in autonomous vehicles versus algorithmic decision-making in criminal justice, financial markets, or social media. In short, algorithmic accountability practices will have to fit the social contexts in which they operate.

When it comes to the forums for algorithmic accountability, perhaps the largest forum is all of those who are or might be affected by algorithmic decision-making. Generally, forums of this scope are mediated through or represented by more specific groups such as the press and media, professional groups of experts, and special interest and activist groups. Organizations like the AI Now Institute and the journal *AI and Society* have already taken up the role and are serving as forums for algorithmic accountability; so have professional organizations such as ACM. Shneiderman has called for oversight and retrospective analysis of disasters involving algorithms and this would create an important forum for addressing the safety of algorithms.[40] More informal and formal forums are needed.

Perhaps, the most striking feature of the current state of algorithmic accountability is the lack of norms for algorithm design and use. Remember that in the case of bridge collapses, the investigation into the cause of failure relied heavily on the norms for bridge construction, maintenance, and use. Norms for the design of algorithms and for their use in decision-making are

[40] Ben Shneiderman, "Opinion: The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight," *Proceedings of the National Academy of Sciences* 113, no. 48 (2016): 13538–13540.

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

direly needed so that actors can be held to account. The need for norms was implicit in Cumming's analysis of the decision points in algorithmic design mentioned earlier. That is, in pointing to the subjectivity in algorithmic design, Cummings effectively pointed to the lack of norms with regard to how algorithm designers select data sets, modify the data sets to make them compatible, decide when the model is accurate, choose which features of the output are significant, and assign weights to features.[41]

Yes, there are informal norms for some of these decision points. That is, designers and researchers may, for example, transmit norms when they train students in how to develop algorithms and when to consider them valid for use in a particular context. And, there may be informal and implicit norms that are transmitted in research literature and professional activities. Nevertheless, such norms, to the extent that they exist at all, are not well articulated or uniformly promulgated; they are not in a form that can be used for public accountability.

Algorithm use is in need of even more attention. Organizational and individual actors seem to adopt the use of algorithms as they see fit, and how those algorithms work is often a black box to the individuals who use them. Yes, in certain domains such as insurance and banking, regulatory mandates constrain the design of algorithms, but such regulatory requirements may not have been developed with an eye to algorithmic implementation. As mentioned earlier, the GDPR requires users to provide explanations to those affected by their algorithm-based decisions; however, even this requirement does not specify when or how algorithms can be used, only that they must be explained to those who are negatively affected. Importantly, recent scholarship pushes for more public attention to whether algorithms should be used at all in certain contexts and calls for more democratic input into the parameters of algorithms.[42] This kind of public attention is sorely needed.

Norms for algorithmic design and use are essential if algorithmic accountability is to be firmly established. I leave aside the matters of where and how algorithmic accountability should be performed and what sort of sanctions there should be for bad actors; these matters depend on the norms that are developed as the most effective for constraining the power of algorithmic decision-making, and on the actors that are identified as accountable.

## V. Conclusion

Returning now to where we began with AI algorithms (which seemed to be the toughest case for accountability) and the endeavor to make algorithms more transparent as a strategy for accountability, the analysis provided here points away from algorithms—be they AI or not—as the key to

[41] Cummings, "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings."
[42] Pasquale, "The Second Wave of Algorithmic Accountability."

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press

algorithmic accountability. The analysis shows that in order to establish algorithmic accountability, attention should be focused on the human actors that design and use algorithms and the norms that should direct their behavior. Algorithmic accountability is still in the making, and the endeavor to make it something effective and significant requires developing a set of practices that have at their heart the idea that actors have an obligation to explain their behavior to forums. This in turn requires norms specifying when actors can be called to account, for what, what counts as an adequate explanation, and what consequences can follow if the explanation is not given or is inadequate. Although making algorithms transparent can contribute to this, currently the discourse is in need of much more discussion of who should be held accountable to whom, for what, and how.

*Applied Ethics; Science, Technology, and Society, University of Virginia, USA*

https://doi.org/10.1017/S0265052522000073 Published online by Cambridge University Press