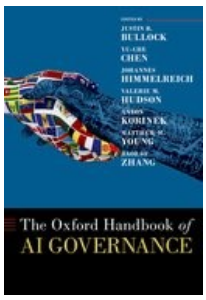

The Oxford Handbook of AI Governance

Justin B. Bullock (ed.) et al.

https://doi.org/10.1093/oxfordhb/9780197579329.001.0001

Published: 2022 **Online ISBN:** 9780197579350 **Print ISBN:** 9780197579329

CHAPTER

8 The Concept of Accountability in AI Ethics and Governance

Theodore M. Lechterman

https://doi.org/10.1093/oxfordhb/9780197579329.013.10 Pages 164–182

Published: 19 December 2022

Abstract

Calls to hold artificial intelligence to account are intensifying. Activists and researchers alike warn of an "accountability gap" or even a "crisis of accountability" in AI. Meanwhile, several prominent scholars maintain that accountability holds the key to governing AI. But usage of the term varies widely in discussions of AI ethics and governance. This chapter begins by disambiguating some different senses and dimensions of accountability, distinguishing it from neighboring concepts, and identifying sources of confusion. It proceeds to explore the idea that AI operates within an accountability gap arising from technical features of AI as well as the social context in which it is deployed. The chapter also evaluates various proposals for closing this gap. It concludes that the role of accountability in AI ethics and governance is vital but also more limited than some suggest. Accountability's primary job description is to verify compliance with substantive normative principles —once those principles are settled. Theories of accountability cannot ultimately tell us what substantive standards to account *for*, especially when norms are contested or still emerging. Nonetheless, formal mechanisms of accountability provide a way of diagnosing and discouraging egregious wrongdoing even in the absence of normative agreement. Providing accounts can also be an important first step toward the development of more comprehensive regulatory standards for AI.

Keywords: artificial intelligence, accountability, responsibility, democracy, AI ethics, AI governance, conceptual analysis

Subject: Political Institutions, Politics

Series: Oxford Handbooks

Collection: Oxford Handbooks Online

Calls to hold artificial intelligence to account are intensifying. Activists and researchers alike warn of an "accountability gap" or even a "crisis of accountability" in AI.[1] Meanwhile, several prominent scholars maintain that accountability holds the key to governing AI.[2] Progress on accountability, they contend, will unlock solutions to numerous other challenges that AI poses, including algorithmic bias, the unintelligibility of algorithmic decisions, and the harmful consequences of certain AI applications.

Appeals to accountability among AI commentators reflect a general trend in public discourse that both lionizes the concept and struggles to specify what it means. Historians of accountability note that the term originates in practices of financial record-keeping and only entered mainstream usage in the late twentieth century.[3] It has since become an ever-expanding concept, used both for narrow purposes and as a catch-all term for normative desirability.[4] This conceptual fuzziness is on full display in AI debates, where no two commentators seem to use the term in precisely the same way. Some scholars treat accountability as a kind of master virtue, using accountability as more or less synonymous with moral justifiability.[5] According to this perspective, AI is accountable when all its features are justifiable to all concerned. Others assign accountability a far more limited role, such as to verify that algorithms comply with existing legal standards or that aspects of system performance are traceable.[6] Some understand accountability as a mechanism for regulating professional roles and organizational relationships;[7] others suggest that accountability is a basic component of moral responsibility that exists independently of institutional practices.[8] Some presume that accountability is a quality of those who design and deploy AI systems,[9] while others treat accountability as a quality of the systems themselves.[10] Because these conceptual disagreements are rarely made explicit, participants in debates about AI accountability often talk past each other.

This chapter begins by disambiguating some different senses and dimensions of accountability, distinguishing it from neighboring concepts, and identifying sources of confusion. It proceeds to explore the idea that AI operates within an accountability gap arising from technical features of AI as well as the
p. 165 social context in which it is deployed. The chapter also ↳ evaluates various proposals for closing this gap. I conclude that the role of accountability in AI ethics and governance is vital but also more limited than some suggest. Accountability's primary job description is to verify compliance with substantive normative principles—once those principles are settled. Theories of accountability cannot ultimately tell us what substantive standards to account *for*, especially when norms are contested or still emerging. Nonetheless, formal mechanisms of accountability provide a way of diagnosing and discouraging egregious wrongdoing even in the absence of normative agreement. Providing accounts can also be an important first step toward the development of more comprehensive regulatory standards for AI.

Different Meanings of Accountability

Analyses tend to agree that accountability is a relational concept with multiple terms.[11] It involves some agent accounting *to* some other agent *for* some state of affairs *according to* some normative standard. A person can be accountable *to* their neighbors *for* noise pollution *according to* local ordinances and commonsense norms of decency. An employee can be accountable *to* an employer *for* the employee's work output *according to* the standards specified by their contract. But as these examples indicate, accountability can apply in wider and narrower senses, with different corresponding standards. A disrespectful neighbor transgresses general moral and legal standards, standards that apply regardless of any contractual agreement. Accountability can be understood, in this first instance, as a dimension of moral responsibility concerned with identifying the causes of states of affairs and assigning praise and blame. In such a case, we seek to the determine the source of the noise, assess whether wrongdoing has occurred, and apply any appropriate sanctions or demands for redress. By contrast, an employee's performance may have little to do with injured parties or independent standards of rightness or wrongness; holding the employee accountable involves the employer assessing the work against the terms of their contract. Accountability in this second instance is a more context-dependent quality. It arises within social practices and relationships where power is delegated from one party to another. Waldron refers to the first sense of accountability as *forensic accountability* and the latter as *agent accountability*.[12] I explore each of these senses of accountability in turn, while also registering the possibility of a third sense of accountability, *accountability-as-a-virtue*.

Accountability in the forensic sense is backward looking and relates closely to responsibility. Theories of responsibility seek to explain how individuals can be connected to their actions and the consequences of their actions in ways that make it appropriate to praise or blame them.[13] In common speech, responsibility and accountability are sometimes used interchangeably. But several philosophers understand accountability more specifically as a component of responsibility. On one prominent view, accountability refers to the conditions under which it is appropriate or fair to *hold someone responsible* for states of affairs.[14] Holding someone responsible is not the same as *believing that someone is responsible*. A victim of sustained psychological trauma may have their moral faculties blunted, leading them to commit a crime. We may believe this person to be responsible for the crime, in the sense that the act was theirs and they performed it with ill intentions. Still, we may believe the person not entirely accountable for the crime, because fully
p. 166 blaming or ↳ sanctioning them would be unfair. To be accountable, according to this understanding, is to be susceptible to a demand for justification, to be expected to provide answers or render an account of what happened and why. It may also involve susceptibility to sanction if the justification comes up short.[15]

Importantly, this perspective holds that responsibility is a prerequisite for accountability: one cannot be accountable for a condition unless one is also responsible for that condition. Discussion of accountability and responsibility in the context of AI has tended to understand this relationship differently. Artificial agents can be accountable—i.e., can be susceptible to demands for justification or sanction—without necessarily being responsible or blameworthy. Floridi and Sanders argue that AI (at least in current and near-term forms) cannot be responsible for wrongdoing.[16] Much like nonhuman animals, AI cannot be responsible because it does not have the relevant intentional states. But AI can be accountable for wrongdoing because it can be sanctioned by modifying or deleting it. Similarly, in her classic study of accountability in a computerized society, Nissenbaum holds that responsibility is sufficient but not necessary to ground a demand for accountability.[17] An agent's being responsible for wrongdoing generates a reason to hold that agent accountable. But one can be accountable for a state of affairs without being responsible for it, such as if the state of affairs was caused by one's subordinate, one's pet animal, or one's technological artifact.

Discussions of accountability in the forensic sense treat accountability as a property to be attributed retrospectively in connection with discrete events. When a technological artifact is involved in some bad

event, we seek to determine who or what is accountable for this, and to treat them accordingly. However, discussions of accountability as a dimension of responsibility also suggest that accountability might be understood as a virtue to be cultivated proactively.[18] An accountable individual, according to this understanding, is one who is robustly disposed to answer for their conduct, to welcome scrutiny of their decisions, and to take responsibility for harms. Likewise, an accountable agent or system is one that reliably welcomes input and oversight from relevant stakeholders, has the right features in place to ensure compliance with relevant standards, and fully acknowledges and rectifies its failures. *Accountability-as-a-virtue* is thus something one can display in greater or lesser quantities.

This way of conceiving accountability resonates with calls in popular discourse for AI—and those who design and deploy it—to be "more accountable." AI and those who deploy it often lack the qualities that enable interested parties to enjoy sufficient input, oversight, or redress. Accountability-as-a-virtue may also help to explain some of the conceptual confusion regarding accountability, as it represents a near antonym of forensic accountability. To be accountable in the forensic sense is generally a negative quality associated with blame and sanctions. But to be accountable in the virtuous sense is a positive quality associated with praise and rewards.

In addition to its usage in moral appraisal and legal investigation, accountability is often described as a more context-dependent quality tied to specific social practices.[19] Practices of accountability involve principal–agent relationships, where one party (the principal) delegates certain tasks or powers to another (the agent) and then monitors performance. The agent owes the principal accounts of this performance according to the terms specified by their relationship, which may be more or less explicit. This *agent accountability*, as Waldron terms it, is the dominant form of accountability within organizations,
p. 167 governments, and professional relationships.[20] It is also one way of characterizing the ↳ relationship between citizens and public officials. Some go so far as to claim that this form of accountability is the "essence" of democracy, as it provides a way for those subjected to coercive power to constrain it.[21]

Problems arise when participants in an accountability relationship implicitly disagree about which model of accountability applies to a given situation. A helpful illustration is the accountability discourse around multilateral organizations. Grant and Keohane report that the World Bank is remarkably accountable to the governments that authorize it but remarkably unaccountable to those affected by its decisions.[22] According to the terminology suggested earlier, defenders of the Bank's accountability implicitly draw upon notions of agent accountability to assess the Bank's accounting to its principals, while critics draw on notions of forensic accountability to assess the Bank's treatment of other stakeholders. Because participants in these debates do not specify which sense of accountability they mean, productive deliberation stalls and tensions escalate.

As this example suggests, complex human societies can have various and overlapping practices of accountability, and confusion often arises over who has standing to demand an account, from whom, and for what. Thus, we can speak of the accountability of AI to its operators or creators, to its users or subjects, to lawmakers or regulators, and to society at large. We can speak of the accountability of AI designers and developers to their superiors, to the law, to industry standards, and to independent moral principles. AI may be perfectly accountable in one dimension but dramatically unaccountable in another dimension. It is not often clear which relationship applies or how subjects of accountability should prioritize amongst competing relationships.

Despite these challenges, the design and deployment of AI systems occur within a dense thicket of formal and informal accountability mechanisms, mechanisms that seek to facilitate the recording, reporting, evaluation, and sanctioning of decisions and activities. Generic accountability mechanisms in modern societies include legislation and law enforcement, the judiciary, government commissions, elections, auditors, whistleblowers, watchdogs, the press, certification standards, professional norms, compliance

departments, and market forces, to name only a few. Each of these elements has a role in preventing transgressions of normative standards, diagnosing these transgressions, or sanctioning transgressions. As discussed below, accountability mechanisms also include a variety of tools proposed for AI specifically, such as transparency and explainability techniques, verification protocols, keeping humans in or on "the loop," and algorithmic impact assessments.

The AI Accountability Gap

The foregoing discussion suggests that two primary conditions need to be in place for accountability to be achieved. First, participants in an accountability relationship must have some basic agreement on the terms: who owes an account to whom for what and according to what standards. Second, subjects of accountability demands must be able to provide accounts according to these terms. Talk of accountability gaps, I propose, reflects a systemic problem with satisfying one or both conditions. A gap may emerge when participants disagree about whether they share an accountability relationship or what its terms are. A gap
p. 168 may also emerge when participants agree on the terms of the relationship but systematically ↳ fail to uphold them for one reason or another. Several features of AI and its social context give rise to accountability gaps. These include (but are not limited to) the distribution of agency across humans and between humans and machines, the opacity and unintelligibility of algorithmic processes, and the persistence of moral and regulatory disagreement.

Distributed agency

One specific limitation to AI's accountability is the way this technology involves distributing agency across numerous human and nonhuman parties. Obviously enough, AI systems may involve the delegation of power from humans to machines. They also typically involve contributions from countless different parties, both human and nonhuman alike. I take up challenges with these features in turn.

The delegation of tasks by humans to autonomous machines involves relinquishing some degree of human control over outcomes. What makes AI novel and valuable is that it provides ways of thinking and acting without human direction and in ways that may be unforeseen by humans. An autonomous vehicle may take us to our destination on a route we never expected; an autonomous weapons system may identify a threat that its human colleagues never considered; AI may diagnose diseases, identify celestial objects, and predict weather all more accurately and more quickly than humans relying on traditional methods. These features are welcome when AI operates in ways that are consistent with human aims and interests, to optimize resource distribution, unravel scientific mysteries, and automate laborious tasks. But precisely because AI is to some extent independent from human understanding and control, it risks acting in ways that are unaccountable to its designers or operators and inconsistent with human aims and interests.

Recent criticism of AI's biased treatment of decision subjects or harms to society reveal that AI's accountability obligations are not limited to its designers and operators. Those who suffer adverse treatment from AI are entitled to demand an account and seek redress. The absence of avenues for appeal and redress of adverse algorithmic decisions is a glaring source of injustice. However, as some observers note, it remains pivotally important that AI be accountable to its designers and operators in the first instance.[23] If AI is not accountable to its designers and operators, it cannot be realistically accountable to anyone else.

This risk of an accountability gap between AI and its human overseers becomes graver the more advanced AI becomes, as there is the potential for AI to reach a level of sophistication where it begins to prioritize its own survival at the expense of human interests.[24] These prospects are somewhat remote, but they could

very well be catastrophic. The risk of autonomous action also becomes grave when AI is used for high-stakes and irreversible applications, such as the exercise of lethal force. Faced with an opportunity to win a war, AI-directed weapons might raze an enemy's cities or eviscerate their own side's human soldiers caught in the crossfire.[25]

AI accountability also faces the problem of "many hands," the notion that AI decisions are ultimately the product of a vast number of different contributions from human and nonhuman agents alike.[26] Algorithmic systems often draw upon third-party datasets and a variety of third-party software elements, both proprietary and open source. The provenance and qualities of these elements may be unknown. Numerous individuals contribute to the collection and classification of data. Numerous further individuals contribute
p. 169 to the design, ↳ testing, and deployment of models, which can be recombined and repackaged over time. The introduction of autonomous operations at various points throughout this sequence further obscures lines of attributability. In some cases, the problem of many hands is a problem because participants failed to record their specific contributions, and the problem could be reduced by requiring better record-keeping. In other cases, the number of operations in a causal chain may be so extensive or so convoluted that it is practically impossible to disentangle individual contributions to final outcomes. Of course, the problem of many hands is not specific to AI. Virtually every product in a modern economy arises from a complex chain of events and contributions before it is consumed. But AI-based products may foster this problem to a greater extent than others, owing to their particular complexity and the fact that some of these hands are not human.

Opacity and unintelligibility

Further accountability risks come from the facts that AI processes are often opaque to human observers, and even when they are more transparent, their decisions are often unintelligible to humans. AI systems may be based on faulty or biased data, they may contain errors in code, and they made encode controversial judgments of their designers. But those who interact with AI systems may not fully understand their purposes, how they work, or what factors they consider when making individual decisions. The problems arise not only from the scale and complexity of AI systems, but also from the proprietary nature of many components. In 2016, a civil society organization exposed that U.S. judges were using a biased algorithmic tool for making sentencing decisions, a tool whose criteria they did not understand—at least partly because the methodology was proprietary.[27]

A related problem arises in the interpretability of algorithmic decisions by those affected by them. Even when the source code and underlying data are available for scrutiny, the rationales for decisions may be difficult to interpret for experts and laypersons alike. The subject of an adverse decision by an algorithm may have little basis for assessing whether the process treated their case appropriately. The unintelligibility of algorithmic decisions inhibits the giving and receiving of accounts.

Moral and regulatory disagreement

Other things equal, accountability is more likely to be achieved in situations where there is already widespread agreement about normative standards—about what agents are accountable *for*. Consider some contrasts. Commonsense morality provides us with a basic shared understanding of the norms of friendship, which in turn allows us to hold friends to account when they fail to uphold these norms. Tort law and environmental regulations provide specific standards for negligence and pollution levels. Victims of a chemical plant disaster may hold the parties responsible for this disaster to account. However, especially when it comes to emerging technologies like AI, standards of harm and wrongdoing are often immature, unclear, or controversial. People disagree profoundly about the general ethical principles with which AI should comply. For instance, is it permissible for AI to reproduce inequalities in underlying conditions but not intensify them? Or should it be required to counteract these ↳ background inequalities in some way?
p. 170
Should AI seek to nudge users toward complying with certain ideals of wellbeing, or should it err on respecting the liberty of its subjects?

Disagreements about the ethics of AI build upon more general disagreement about the nature of specific values like liberty and equality. They also build upon longstanding disagreements in normative ethics concerning how to appraise rightness or wrongness in general. When we hold AI accountable, should we take primary concern with intentions, actions, or results?[28] A credit-rating algorithm might be designed with the beneficent intentions of expanding access to credit, reducing capricious judgments by loan officers, and reducing loan default rates. Once deployed, it may in fact achieve these results. Despite these good intentions and results, it may also treat certain people unfairly. Different theories and different people place different weight on the significance of intentions, actions, and results in judging rightness and wrongness. When there is widespread disagreement about the standards to account *for*, generic calls for greater accountability appear to lack a clear target. In such cases, I suggest, calls for accountability might be understood as prompts for clarifying the normative standards that are prerequisites for successful accountability practices.

In addition to disagreement about which moral standards apply to AI, there is also tremendous disagreement over which regulatory standards apply to AI. Laws and industry conventions are still embryonic and competing for dominance. National governments and intergovernmental organizations have proposed many regulatory frameworks but so far passed little legislation. Seemingly, every professional association, standard-setting organization, and advocacy group is hawking a different list of principles, guidelines, and tools for regulating AI.[29] These efforts indicate broad agreement on the significance of the ethical challenges that AI poses. And many of these efforts reflect similar themes. But unless or until there is consolidation of competing terms, principles, and protocols, the accountability of AI is likely to suffer. Paradoxically, an abundance of competing standards can reduce accountability overall by inviting confusion and creating opportunities for actors to pick and choose the standards that burden them least.[30]

Closing the AI Accountability Gap

There are many proposals for closing aspects of AI's accountability gap. Some are more promising than others. Less promising proposals include attempts to ban broad categories of AI, initiatives to regulate AI as a general class, demands to make AI transparent, and proposals to make technology professionals the primary guardians of AI ethics. Alternatives include contextually sensitive regulatory approaches that appreciate differences in technological functions and domains of application; traceability and verification techniques; and a division of labor that expects professionals to flag ethical dilemmas without having the exclusive authority for adjudicating them.

Moratoria

One way to close an AI accountability gap is to eliminate its very possibility. Some suggest banning AI
p. 171 altogether or banning AI in entire domains of application, such as defense, ↳ healthcare, or transportation.[31] Certain governments, including those of San Francisco, Oakland, Seattle, and Morocco, have enacted temporary bans on facial recognition until appropriate regulations can be devised.[32] Although narrowly crafted bans may indeed be warranted in cases like these, categorical bans face particular objections. One is that they may exceed their justifiable scope. Certain uses of AI in military applications or healthcare may be far less risky, or far more susceptible to accountability, than others. Automating target selection and choice of means is one thing. Using AI to assist human decision-makers about these things is another. And using AI for non-combat purposes, such as to optimize logistics or triage in humanitarian crises, is another thing altogether. Similarly, automating medical diagnoses and treatment decisions for life-threatening conditions may indeed create unacceptable risks. But these risks may not arise to the same extent when using AI to assist doctors in low-stakes diagnostic questions.[33]

Proposals to ban technologies must also be sensitive to the possibility of prisoners' dilemmas that may counteract a ban's intended effects. If Country A bans AI for military use, Country B gains a strategic advantage by continuing to develop AI for military use. If both countries agree to ban AI for military use, fear that either one may covertly continue development provides an incentive for each to continue development in secret. And even if effective monitoring mechanisms can make these commitments credible, there is always the possibility of a black market of non-state actors developing killer robots for the highest bidder.[34] If autonomous weapons are likely to be developed no matter what Country A does, banning all research and development may be self-defeating as it would put Country A at greater risk from attacks from other countries' autonomous weapons.

A third problem is that bans necessarily involve foregoing the potential benefits of AI, which can be tremendous. Automobile accidents kill 36,000 people in the United States each year, driven significantly by speeding, distraction, and driving under the influence.[35] Autonomous vehicles, which do not suffer from these problems, are expected to dramatically reduce road deaths, even as they may introduce or exacerbate other problems. Regulatory discussions that endorse a "precautionary principle" can fail to appreciate the opportunity costs of preserving the status quo.[36] Victims of traditional car crashes who would otherwise survive should autonomous vehicles be introduced have a powerful objection to postponing or preventing the deployment of self-driving cars.

Ironically, certain problems that moratoria aim to fix might be reduced by allowing a system to iterate and dynamically improve in large-scale applications. Thus, a diagnostic AI trained on biased data becomes dramatically more accurate the more patient data it receives. Sometimes this data can only be made available by releasing the product publicly. One difficulty here, of course, is the fairness to first-generation users of new technology, who must bear the consequences of less reliable products. But this problem is hardly unique to AI, and solutions to it have a long history in public health, for instance.

Some, but certainly not all, of the concerns animating calls to ban applications of AI stem from fully automated uses in which humans are not directly involved in the decision-making process or absent entirely. The now-familiar typology distinguishes between having humans "in the loop" (receiving advice from AI but responsible for determining whether and how to act on that advice), "on the loop" (where AI implements its own decisions while humans monitor and intervene if necessary), and "out of the loop" (where humans are not actively involved in deciding or monitoring).[37] In some cases, the accountability gap
p. 172 shrinks by keeping humans more closely involved and using AI primarily to augment human ↳ intelligence rather than replace it completely.[38] As discussed further below, we have reason to worry about whether the humans in the loop are themselves the appropriate decision-makers, as those who design or operate AI and

those who suffer the consequences of AI decisions are often not identical. But this issue is in principle separate from the question of human control itself.

Regulatory approaches: All-purpose and contextual

The steady stream of alarming mistakes and doomsday scenarios reveal the limits of patchwork regulatory standards and prompt increasing calls to regulate AI as a general class. Tutt, for instance, proposes a new federal agency modeled after the U.S. Food and Drug Administration to oversee the testing and approval of algorithms.[39] Such proposals would require imposing a common set of normative standards, technical criteria, and/or reporting requirements on all forms of AI. This would certainly make AI more formally accountable, but it would come with significant tradeoffs. AI is not monolithic and varies tremendously in its moral risks. Calls to regulate AI as a general class can fail to appreciate that many uses of AI are largely privately regarding and contain limited risks of harm. Consider AI applications for composing music.[40] Such technology might introduce or intensify disputes over intellectual property, but it raises no obvious threats to health, safety, or equality, and the case for granting the state additional oversight here appears relatively weak. Undifferentiated demands for public accountability can infringe on behavior that is more or less benign and privately concerned.

Although AI for music composition and AI for judicial sentencing clearly occupy opposite poles on the private-public scale, many applications of AI occupy a more nebulous intermediate area. Recommender algorithms on search engines and social media platforms are cases in point. Search engines and social media platforms are private corporations, but they can come to monopolize the flow of information with dramatic effects on public discourse and political stability.

A more promising approach to regulation would take account of various contextual factors, such as the domain of operation, the kinds of agents involved, asymmetries in information and power, and the different interests at stake. Different standards might apply based on whether subjection to the decisions is voluntary or nonvoluntary, whether the decisions are high-stakes or low-stakes, whether the risks of externalities are high or low, the degree of human oversight, the degree of competition, and so on. This idea has much in common with Nissenbaum's noted theory of privacy as "contextual integrity," a view holding that privacy is not an independent value but one that demands different things in different settings.[41]

Transparency and explainability

Talk of closing the accountability gap often appeals to principles of transparency and explainability.[42] Improving the transparency and explainability of AI is often claimed to be a major component of improving accountability, as we seem unable to determine whether AI complies with the reasons that apply to it if we
p. 173 cannot understand what it decides and ↳ why. There is certainly a role for improvements in both qualities in making AI more accountable. But singular focus on either element leads to certain traps. As Kroll has argued, transparency is often neither desirable nor sufficient for making AI accountable.[43] It is not desirable in uses that require the protection of data subject privacy, trade secrets, or national security. It is not sufficient in most cases, as merely being able to view the data set or code of an algorithm is hardly a guarantee of making sense of it. When Reddit released code to the public indicating how its content moderation algorithm works, prominent computer scientists could not agree on how to interpret it.[44]

Demands for transparency often appear rooted in the implicit belief that transparency conduces to explainability or interpretability. If we can view the data or the code, this thinking goes, we are more likely to understand the algorithm's decisions. Although there continues to be interesting research and experimentation on improving the explainability of algorithmic decisions, to a certain extent the search for explainability is chimerical. The most advanced forms of AI are not programmed by humans but rather

result from deep learning processes, which automatically create and adjust innumerable settings in response to training data. What these settings mean and why they were selected may be virtually unknowable. The more complex AI becomes, the harder its processes are to understand, and efforts to reverse-engineer them come with their own biases and limitations.[45] In situations where precise explanation is essential to the justification of a decision, as in criminal sentencing, it may be wiser to regulate the use of AI than to demand explainability from AI. Indeed, some propose that in high-stakes or public administration settings, the use of "black box" AI models is simply impermissible.[46] Decision-makers in these settings may only permissibly rely upon algorithmic tools that are interpretable by design and sufficiently well understood by their designers and operators.

A variety of alternative methods have been proposed to monitor the integrity and reliability of algorithmic systems in the absence of transparency and explainability. These include the banal but often overlooked methods of robust documentation and record-keeping,[47] clear divisions of responsibility during the development process, publicizing and following a set of standard operating procedures,[48] and different visualization and reporting methods to track and communicate the qualities of an algorithm, such as dashboards and "nutrition labels." Of particular interest to many are algorithmic impact assessments, which seek to forecast, declare, and offer mitigation strategies for potential adverse effects of a given AI application.[49] More technical tools include software verification techniques that check whether software matches it specifications and the use of cryptography to authenticate features and performance.[50] These methods cannot make AI fully explainable, but they can provide grounds for greater confidence in the results of AI decisions in certain cases.

Duty of care

Another proposed solution to the AI accountability gap involves tasking those who design and deploy AI with a duty of care to mitigate ethical risks of AI systems.[51] Many risks of AI can indeed be mitigated by heightened sensitivity of designers and operators to ethical issues. Greater awareness of structural injustice and the kinds of biases that may lurk in training data might be enough to prevent certain horrendous mistakes like the release of facial recognition products that classify Black faces as gorillas.[52]

p. 174 However, a duty of care can be easily abused. Many ethical issues are too complex to be solved without more advanced expertise, and the ethical hubris of many technology professionals is already legendary.[53] Inviting professionals to take responsibility for ethically safeguarding their products can be a recipe for well-meaning mistakes, motivated reasoning, or encoding parochial value judgments into software. Many ethical issues arguably exceed the authority of technology professionals to resolve on their own. Plenty of these demand input from affected communities and a fair process of public deliberation. Overzealous exercise of the duty of care may invite criticism of paternalism or technocracy.

Some suggest that objections to the private governance of AI can be mitigated by limiting the range of eligible justifications for AI designs and outcomes. The criteria we use to appraise AI must operate within the bounds of "public reason"—reasons that any and every citizen could be expected to endorse.[54] This solution may certainly help to screen out the most parochial or controversial justifications, such as those rooted in narrow conceptions of human flourishing or faulty logic. But much, if not most, of the current disagreement in AI ethics already operates within the realm of public reason and appeals to public reason are of little help in resolving these debates.

An alternative approach to a duty of care is to train technology professionals on identifying and flagging ethical issues to be adjudicated by others. Designers and operators are the first line of defense in detecting potential harms from AI. With training, they may become attuned to noting the presence of controversial assumptions, disparate impacts, and value trade-offs. But deeper sensitivity to these ethical risks and

appropriate ways of resolving them may profit from interdisciplinary collaboration between computing professionals and experts from academia and civil society. It is also a ripe opportunity for experimentation with new forms of civic engagement that allow input on technical questions by those affected by them.[55]

AI as an Accountability Instrument

The foregoing discussion has explored some of the challenges of ensuring that AI and those who design and apply it are accountable. However, it also pays to consider how AI might both erode and improve the accountability of conventional entities. AI can enable malicious actors and systems to evade accountability. It can also serve as an instrument for facilitating the accountability of humans and institutions.

Although not an instance of AI, blockchain is an adjacent form of digital technology that exemplifies this duality. A blockchain is a distributed ledger that uses cryptography to store value, facilitate exchanges, and verify transactions. Blockchain has applications in the verification of identities, the storage of digital assets, the assurance of contract fulfillment, and the security of voting systems. It creates strong mutual accountability by reducing reliance on individual trust or third-party institutions like governments, lawyers, and banks. Blockchain is most well-known for its use in cryptocurrency, decentralized media of exchange that are not authorized or controlled by central banks. Cryptocurrency is especially helpful to people and places ill-served by fiat currencies, where banking services may be inaccessible, dysfunctional, or discriminatory. But cryptocurrency's ungovernability creates a double-edged sword. An ungovernable currency becomes the medium of choice for illicit ↳ transactions.[56] Furthermore, while major players can influence elements of the market, ordinary cryptocurrency holders have no way of holding the system to account for adverse conditions.[57]

p. 175

The ways that AI can facilitate state surveillance and law enforcement are increasingly apparent in the forms of predictive policing, facial recognition, and judicial sentencing algorithms. Naturally, AI has numerous beneficial applications in government and can promote decisions that are more just and legitimate. In theory, decisions driven by rigorous data analysis can result in outcomes that are more efficient, consistent, fair, and accurate. Given optimistic assumptions about its ability to overcome challenges of bias and opacity, AI may even improve government accountability by reducing reliance on human discretion. But by expanding the power of states for surveilling subjects, controlling populations, and quashing dissent, AI also supplies states with powerful means for evading accountability.

As Danaher discusses (albeit skeptically), AI can also be part of the solution to state oppression by powering "sousveillance" methods that hold powerful actors to account.[58] Sousveillance refers to watching from below, and it is exemplified by efforts to film police misconduct on smartphones. AI in the hands of citizens and civil society groups may facilitate sousveillance by enabling the powerless to analyze data for signs of misconduct. This might take the form of journalists pursuing freedom of information requests, criminal justice advocates analyzing forensic evidence for signs of false convictions, or human rights activists tracking abuse through posts on social media. The tools of sousveillance also extend to consumer protection, as with applications that use bots to challenge bank fees, product malfunctions, and price gouging.

Conclusion: Accountability's Job Specification

We should be wary of placing too much faith in accountability as such. Greater accountability does not necessarily lead to greater justice. Functionaries who faithfully comply with the dictates of a genocidal regime are eminently accountable, in many respects. Software engineers might be perfectly accountable to their superiors, whose aim is to maximize profits at any social cost.

Some suggest that accountability is the "essence" of democracy, as it provides a constraint on unchecked power.[59] This position, however, also finds support with certain skeptics of democracy, who have sought to limit participation in politics to periodic opportunities to check abuses of power without opportunities to exercise or influence power in the first place.[60] For proponents of a more demanding view of the democratic ideal, accountability is better understood as but one feature of democratic legitimacy: namely, a condition on policy outcomes. For these perspectives, democratic legitimacy also requires conditions on policy inputs, such as collective self-determination, political equality, and deliberative decision-making.[61]

Debating accountability and its mechanisms can also distract us from fundamental questions about substantive normative standards. If we do not adequately address the question of what principles should
p. 176 regulate the design and use of AI and under what ↳ conditions, debate about whether and how AI can be accountable to those principles seems to lose much of its point.

Despite accountability's limitations, however, the claim that accountability holds the key of AI ethics and governance is worth taking seriously. Especially when there is disagreement about substantive standards, accountability mechanisms may play an essential role in the discovery of problems and the search for more lasting solutions.[62] Procedural regularity, documentation, and impact assessments enable accounts to be given. There may be disagreement about the normative standards that apply to these accounts, but having the accounts is a critical step toward diagnosing problems, refining standards, and sanctioning failures. While accountability may not be all that democracy demands, institutional, organizational, and technical mechanisms that enable scrutiny of power are absolutely crucial to the protection and realization of democratic ideals.

Moreover, agreement on the details of principles of justice is not necessary for seeking accountability for violations of basic human rights and other obvious harms. There is already widespread agreement about certain fundamental rights and duties, and not all grounds for disagreement are reasonable. Improving the accountability of AI to basic moral standards would leave much work to be done, but it would also constitute a remarkable achievement.

Still, as this chapter has emphasized, the concept of accountability contains many puzzles and remains poorly understood. Improving the accountability of AI may be difficult to achieve without further work to disentangle and narrow disagreement on the concept's different meanings and uses.

Acknowledgments

For extraordinarily helpful comments on earlier drafts, I thank Johannes Himmelreich, Juri Viehoff, Jon Herington, Kate Vredenburgh, Carles Boix, David Danks, audience members at the 2021 Society for the Philosophy of Technology annual conference, and four anonymous readers for Oxford University Press.

Notes

1. Institute for the Future of Work. (2020). Mind the gap: How to fill the equality and AI accountability gap in an automated world. London, October; Raji, Inioluwa Deborah et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona Spain, pp. 33–44, https://doi.org/10.1145/3351095.3372873; Hutchinson, Ben et al. (2020). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *ArXiv:2010.13561 [Cs]*, October 22, http://arxiv.org/abs/2010.13561.

2. Wachter, Sandra, Mittelstadt, Brent, & Floridi, Luciano. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics 2* (6), eaan6080, https://doi.org/10.1126/scirobotics.aan6080; Skelton, Sebastian Klovig. (2019). Accountability is the key to ethical artificial intelligence, experts say. *ComputerWeekly.Com*, December 16, https://www.computerweekly.com/feature/Accountability-is-the-key-to-ethical-artificial-intelligence-experts-say.

p. 177 3. Bovens, Mark et al. (2014). Public accountability. In Mark Bovens, Robert E. Goodin, & Thomas Schillemans (Eds.), *The Oxford handbook of public accountability* (pp. 1–20). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199641253.013.0012.

4. Mulgan, Richard. (2000). "Accountability": An ever-expanding concept? *Public Administration 78* (3), 555–573, https://doi.org/10.1111/1467-9299.00218.

5. See, e.g., Dignum, Virginia. (2020). Responsibility and artificial intelligence. In Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 218). Oxford University Press.

6. Kroll, Joshua A. et al. (2016). Accountable algorithms. *University of Pennsylvania Law Review 165*, 633; Kohli, Nitin, Barreto, Renata, & Kroll, Joshua A. (2018). Translation tutorial: A shared lexicon for research and practice in human-centered software systems. 1st Conference on Fairness, Accountability, and Transparency, New York.

7. Bovens, Mark. (2007). Public accountability. In Ewan Ferlie, Laurence E. Lynn, Jr., & Christopher Pollitt (Eds.), *The Oxford handbook of public administration* (pp. 182–208). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780199226443.003.0009.

8. See, e.g., Watson, Gary. (1996). Two faces of responsibility. *Philosophical Topics 24* (2), 227–248, https://doi.org/10.5840/philtopics199624222; Shoemaker, David. (2011). Attributability, answerability, and accountability: toward a wider theory of moral responsibility. *Ethics 121* (3), 602–632, https://doi.org/10.1086/659003.

9. Nissenbaum, Helen. (1996). Accountability in a computerized society. *Science and Engineering Ethics 2* (1), 25–42, https://doi.org/10.1007/BF02639315.

10. Floridi, Luciano, & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines 14* (3), 349–379, https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

11. Goodin, Robert E. (2003). Democratic accountability: the distinctiveness of the third sector. *European Journal of Sociology 44* (3): 359–396, https://doi.org/10.1017/S0003975603001322; Watson, Gary. (1996). Two faces of responsibility. *Philosophical Topics 24* (2), 227–248, https://doi.org/10.5840/philtopics199624222.

12. Waldron, Jeremy. (2016). Accountability and insolence. In *Political political theory: Essays on institutions* (pp. 167–194). Harvard University Press.

13. Noorman, Merel. (2018). Computing and moral responsibility. In Edward N. Zalta (Ed.), *Stanford encyclopedia of philosophy*, https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.

14. Watson, "Two faces of responsibility."

15. Shoemaker, David. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics 121* (3), 602–632, https://doi.org/10.1086/659003.

16. Floridi & Sanders, "On the morality of artificial agents."

17. Nissenbaum, “Accountability in a computerized society.”

18. Bovens, Mark. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics 33* (5), 946–967, https://doi.org/10.1080/01402382.2010.486119.

19. Goodin, “Democratic accountability”; Bovens, “Public accountability”; Bovens et al., “Public accountability”; Waldron, “Accountability and insolence.”

20. Waldron, “Accountability and insolence.”

21. Bovens, “Public accountability.”

22. Grant, Ruth W., & Keohane, Robert O. (2005). Accountability and abuses of power in world politics. *American Political Science Review 99* (1), 29–43, https://doi.org/10.1017/S0003055405051476.

p. 178 23. Wagner, Ben. (2020). Algorithmic accountability: Towards accountable systems. In Giancarlo Frosio (Ed.), *The Oxford handbook of online intermediary liability* (pp. 678–688). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780198837138.013.35.

24. Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking Press.

25. Asaro, Peter. (2020). Autonomous weapons and the ethics of artificial intelligence. In S. Matthew Liao (Ed.), *Ethics of artificial intelligence* (pp. 212–236). Oxford University Press, https://doi.org/10.1093/oso/9780190905033.003.0008.

26. Nissenbaum, “Accountability in a computerized society.”

27. Noorman, “Computing and moral responsibility.”

28. Goodin, “Democratic accountability.”

29. In 2019, one study counted 84 different initiatives to articulate ethical principles for AI. See Mittelstadt, Brent (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1* (11), 501–507. As of July 2021, a repository at www.aiethicist.org contained hundreds of different governmental and nongovernmental proposals for defining and upholding AI norms.

30. For a study of how the existence of overlapping accountability demands can go awry, see Koppell, Jonathan G. S. (2005). Pathologies of accountability: ICANN and the challenge of “multiple accountabilities disorder”. *Public Administration Review 65* (1), 94–108, https://doi.org/10.1111/j.1540-6210.2005.00434.x.

31. Asaro, “Autonomous weapons and the ethics of artificial intelligence.”

32. Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021). Algorithmic accountability for the public sector, August, https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/, p. 16.

33. Of course, the line between a low-stakes and high-stakes diagnostic question in medicine is often fuzzy, as symptoms of serious illness may often present similarly to symptoms of superficial illness.

34. The concern here is that unlike nuclear weapons and other weapons of mass destruction, the development of autonomous weapons has low barriers to entry and may be far more difficult to monitor. See, e.g., McGinnis, John O. (2010). Accelerating AI. *Northwestern University Law Review 104* (3), 1253–1269.

35. Insurance Institute for Highway Safety/Highway Loss Data Institute. (2021, March). Fatality Facts 2019: Yearly Snapshot, https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot.

36. Sunstein, Cass R. (2003). Beyond the precautionary principle. *University of Pennsylvania Law Review 151* (3), 1003–1058, https://doi.org/10.2307/3312884.

37. For an overview, see Rahwan, Iyad. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology 20* (1), 5–14, https://doi.org/10.1007/s10676-017-9430-8.

38. This is not to say that keeping humans in the loop is a panacea. The tendency of humans to trust too readily in the judgments of machines is a well-known source of cognitive bias. See, e.g., Skitka, Linda J., Mosier, Kathleen, & Burdick,

Mark D. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies 52* (4), 701–717, https://doi.org/10.1006/ijhc.1999.0349.

39. Tutt, Andrew. (2017). An FDA for algorithms. *Administrative Law Review 69* (1), 83–123.

p. 179 40. See, e.g., Fernandez, J. D., & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research 48*, 513–582, https://doi.org/10.1613/jair.3908.

41. Nissenbaum, Helen. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

42. Doshi-Velez, Finale et al. (2019). Accountability of AI under the law: The role of explanation. Preprint. *ArXiv:1711.01134 [Cs.AI]*, December 20, http://arxiv.org/abs/1711.01134.

43. Kroll, Joshua A. (2020). Accountability in computer systems. In Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (179–196). Oxford University Press, https://doi.org/10.1093/oxfordhb/9780190067397.013.10.

44. New, Joshua, & Castro, Daniel. (2018). How policymakers can foster algorithmic accountability. Center for Data Innovation, May 21.

45. Rudin, Cynthia. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence 1* (5), 206–215, https://doi.org/10.1038/s42256-019-0048-x.

46. Ibid. See also, Busuioc, Madalina. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review 81* (5), 834.

47. Raji et al., "Closing the AI accountability gap."

48. Kroll et al., "Accountable algorithms."

49. For critical discussion, see Selbst, Andrew D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology 35* (1), 117–191.

50. Ibid.

51. Nissenbaum, "Accountability in a computerized society."

52. Guynn, Jessica. (2015). Google photos labeled Black people "gorillas". *USA Today*, July 1.

53. Morozov, Evgeny. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.

54. Binns, Reuben. (2018). Algorithmic accountability and public reason. *Philosophy & Technology 31* (4), 543–556, https://doi.org/10.1007/s13347-017-0263-5.

55. Landemore, Hélène. (2020). *Open democracy: Reinventing popular rule for the twenty-first century*. Princeton University Press.

56. Foley, Sean, Karlsen, Jonathan R., & Putniņš, Tālis J. (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies 32* (5), 1798–1853, https://doi.org/10.1093/rfs/hhz015; Kethineni, Sesha, & Cao, Ying. (2020). The rise in popularity of cryptocurrency and associated criminal activity. *International Criminal Justice Review 30* (3), 325–344, https://doi.org/10.1177/1057567719827051.

57. Atzori, Marcella. (2017). Blockchain technology and decentralized governance: Is the state still necessary? *Journal of Governance and Regulation 6* (1), 45–62, https://doi.org/10.22495/jgr_v6_i1_p5.

58. Danaher, John. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology 29* (3), 245–268, https://doi.org/10.1007/s13347-015-0211-1.

59. Bovens, "Public accountability."

60. Schumpeter, Joseph A. (2008). *Capitalism, socialism, and democracy*. 1st ed. Harper Perennial Modern Thought.

61. Christiano, Thomas. (1996). *The rule of the many: Fundamental issues in democratic theory*. Westview Press.

62. Kroll, “Accountability in computer systems.”

References

Ada Lovelace Institute, AI Now Institute, & Open Government Partnership. (2021). Algorithmic accountability for the public sector. August. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.
WorldCat

Atzori, Marcella. (2017). Blockchain technology and decentralized governance: Is the state still necessary? *Journal of Governance and Regulation* 6 (1), 45–62. https://doi.org/10.22495/jgr_v6_i1_p5.
Google Scholar WorldCat

Binns, Reuben. (2018). Algorithmic accountability and public reason. *Philosophy & Technology* 31 (4), 543–556. https://doi.org/10.1007/s13347-017-0263-5.
Google Scholar WorldCat

Bovens, Mark. (2007). Public accountability. In Ewan Ferlie, Laurence E. Lynn, Jr., and Christopher Pollitt (Eds.), *The Oxford handbook of public administration* (182–208). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199226443.003.0009.
Google Scholar Google Preview WorldCat COPAC

Bovens, Mark. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism." *West European Politics* 33 (5), 946–967. https://doi.org/10.1080/01402382.2010.486119.
Google Scholar WorldCat

Bovens, Mark, Goodin, Robert E., & Schillemans, Thomas. (2014). Public accountability. In Mark Bovens, Robert E. Goodin, and Thomas Schillemans (Eds.), *The Oxford handbook of public accountability* (pp. 1–20). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0012.
Google Scholar Google Preview WorldCat COPAC

Busuioc, Madalina. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review* 81 (5), 825–836. https://doi.org/10.1111/puar.13293.
Google Scholar WorldCat

Christiano, Thomas. (1996). *The rule of the many: Fundamental issues in democratic theory*. Westview Press.
Google Scholar Google Preview WorldCat COPAC

Danaher, John. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology* 29 (3), 245–268. https://doi.org/10.1007/s13347-015-0211-1.
Google Scholar WorldCat

Dignum, Virginia. (2020). Responsibility and artificial intelligence. In Markus D. Dubber, Frank Pasquale, and Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 213–231). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.12.
Google Scholar Google Preview WorldCat COPAC

Doshi-Velez, Finale, Kortz, Mason, Budish, Ryan, Bavitz, Chris, Gershman, Sam, O'Brien, David, Scott, Kate, Schieber, Stuart, Waldo, James, Weinberger, David, Weller, Adrian, & Wood, Alexandra. (2019). Accountability of AI under the law: The role of explanation. Preprint. *ArXiv:1711.01134 [Cs.AI]*, December 20. http://arxiv.org/abs/1711.01134.

Fernandez, J.D., & Vico, F.. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research* 48, 513–582. https://doi.org/10.1613/jair.3908.
Google Scholar WorldCat

Floridi, Luciano, & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines* 14 (3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.
Google Scholar WorldCat

Foley, Sean, Karlsen, Jonathan R., & Putniņš, Tālis J. (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies* 32 (5), 1798–1853. https://doi.org/10.1093/rfs/hhz015.
Google Scholar WorldCat

Goodin, Robert E. (2003). Democratic accountability: The distinctiveness of the third sector. *European Journal of Sociology* 44 (3), 359–396. https://doi.org/10.1017/S0003975603001322.
Google Scholar WorldCat

Grant, Ruth W., & Keohane, Robert O. (2005). Accountability and abuses of power in world politics. *American Political Science Review* 99 (1), 29–43. https://doi.org/10.1017/S0003055405051476.
Google Scholar WorldCat

Guynn, Jessica. (2015). Google photos labeled Black people "gorillas". *USA Today*, July 1.

p. 181

Hutchinson, Ben, Smart, Andrew, Hanna, Alex, Denton, Emily, Greer, Christina, Kjartansson, Oddur, Barnes, Parker, & Mitchell, Margaret. (2020). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. *ArXiv:2010.13561 [Cs]*, October 22. http://arxiv.org/abs/2010.13561.

Institute for the Future of Work. (2020, October). Mind the gap: How to fill the equality and AI accountability gap in an automated world. London.

Kethineni, Sesha, & Cao, Ying. (2020). The rise in popularity of cryptocurrency and associated criminal activity. *International Criminal Justice Review* 30 (3), 325–344. https://doi.org/10.1177/1057567719827051.
Google Scholar WorldCat

Kohli, Nitin, Barreto, Renata, & Kroll, Joshua A. (2018). Translation tutorial: A shared lexicon for research and practice in human-centered software systems. *1st Conference on Fairness, Accountability, and Transparency*. February. New York.

Koppell, Jonathan G.S. (2005). Pathologies of accountability: ICANN and the challenge of "multiple accountabilities disorder". *Public Administration Review* 65 (1), 94–108. https://doi.org/10.1111/j.1540-6210.2005.00434.x.
Google Scholar WorldCat

Kroll, Joshua A. (2020). Accountability in computer systems. In Markus D. Dubber, Frank Pasquale, & Sunit Das (Eds.), *The Oxford handbook of ethics of AI* (pp. 179–196). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.10.
Google Scholar Google Preview WorldCat COPAC

Kroll, Joshua A., Barocas, Solon, Felten, Edward W., Reidenberg, Joel R., Robinson, David G., & Yu, Harlan. (2016). Accountable algorithms. *University of Pennsylvania Law Review* 165, 633.
Google Scholar WorldCat

Landemore, Hélène. (2020). *Open democracy: Reinventing popular rule for the twenty-first century*. Princeton University Press.
Google Scholar Google Preview WorldCat COPAC

McGinnis, John O. (2010). Accelerating AI. *Northwestern University Law Review* 104 (3), 1253–1269.
Google Scholar WorldCat

Mittelstadt, Brent. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1 (11), 501–507. https://doi.org/10.1038/s42256-019-0114-4.
Google Scholar WorldCat

Morozov, Evgeny. (2013). *To save everything, click here: The folly of technological solutionism*. Public Affairs.
Google Scholar Google Preview WorldCat COPAC

Mulgan, Richard. (2000). "Accountability": An ever-expanding concept? *Public Administration* 78 (3), 555–573. https://doi.org/10.1111/1467-9299.00218.

Google Scholar WorldCat

New, Joshua, & Castro, Daniel. (2018). How policymakers can foster algorithmic accountability. Center for Data Innovation.

Nissenbaum, Helen. (1996). Accountability in a computerized society. *Science and Engineering Ethics* 2 (1), 25–42. https://doi.org/10.1007/BF02639315.
Google Scholar WorldCat

Nissenbaum, Helen. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
Google Scholar Google Preview WorldCat COPAC

Noorman, Merel. (2020). Computing and moral responsibility. In Edward N. Zalta (Ed.), *Stanford encyclopedia of philosophy*, Spring Edition. https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.
Google Scholar Google Preview WorldCat COPAC

Rahwan, Iyad. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20 (1), 5–14. https://doi.org/10.1007/s10676-017-9430-8.
Google Scholar WorldCat

Raji, Inioluwa Deborah, Smart, Andrew, White, Rebecca N., Mitchell, Margaret, Gebru, Timnit, Hutchinson, Ben, Smith-Loud, Jamila, Theron, Daniel, & Barnes, Parker. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44. ACM, 2020. https://doi.org/10.1145/3351095.3372873.

p. 182 Rudin, Cynthia. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.
Google Scholar WorldCat

Russell, Stuart J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking Press.
Google Scholar Google Preview WorldCat COPAC

Schumpeter, Joseph A. (2008). *Capitalism, socialism, and democracy*. 1st ed. Harper Perennial Modern Thought.
Google Scholar Google Preview WorldCat COPAC

Selbst, Andrew D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology* 35 (1), 117–191.
Google Scholar WorldCat

Shoemaker, David. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 121 (3), 602–632. https://doi.org/10.1086/659003.
Google Scholar WorldCat

Skelton, Sebastian Klovig. (2019). Accountability is the key to ethical artificial intelligence, experts say. *ComputerWeekly.Com*, December 16. https://www.computerweekly.com/feature/Accountability-is-the-key-to-ethical-artificial-intelligence-experts-say.
WorldCat

Skitka, Linda J., Mosier, Kathleen, & Burdick, Mark D. (2000). Accountability and automation bias. *International Journal of Human–Computer Studies* 52 (4), 701–717. https://doi.org/10.1006/ijhc.1999.0349.
Google Scholar WorldCat

Sunstein, Cass R. (2003). Beyond the precautionary principle. *University of Pennsylvania Law Review* 151 (3), 1003–1058. https://doi.org/10.2307/3312884.
Google Scholar WorldCat

Tutt, Andrew. (2017). An FDA for algorithms. *Administrative Law Review* 69 (1), 83–123.

Google Scholar WorldCat

Wachter, Sandra, Mittelstadt, Brent, & Floridi, Luciano. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics* 2 (6), eaan6080. https://doi.org/10.1126/scirobotics.aan6080.
Google Scholar WorldCat

Wagner, Ben. (2020). Algorithmic accountability: Towards accountable systems. In Giancarlo Frosio (Ed.), *The Oxford handbook of online intermediary liability* (pp. 678–688). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198837138.013.35.
Google Scholar Google Preview WorldCat COPAC

Waldron, Jeremy. (2016). Accountability and insolence. In *Political political theory: Essays on institutions* (pp. 167–194). Harvard University Press.
Google Scholar Google Preview WorldCat COPAC

Watson, Gary. (1996). Two faces of responsibility. *Philosophical Topics* 24 (2): 227–248.
https://doi.org/10.5840/philtopics199624222.
Google Scholar WorldCat