Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning

A. Feder Cooper*
Cornell University
Ithaca, NY, USA
afc78@cornell.edu

Emanuel Moss*
Cornell Tech and Data & Society Research Institute
New York, NY, USA
emanuelmoss@cornell.edu

Benjamin Laufer
Cornell Tech
New York, NY, USA
bdl56@cornell.edu

Helen Nissenbaum
Cornell Tech
New York, NY, USA
hn288@cornell.edu

ABSTRACT

In 1996, *Accountability in a Computerized Society* [95] issued a clarion call concerning the erosion of accountability in society due to the ubiquitous delegation of consequential functions to computerized systems. Nissenbaum [95] described four barriers to accountability that computerization presented, which we revisit in relation to the ascendance of data-driven algorithmic systems—i.e., machine learning or artificial intelligence—to uncover new challenges for accountability that these systems present. Nissenbaum's original paper grounded discussion of the barriers in moral philosophy; we bring this analysis together with recent scholarship on relational accountability frameworks and discuss how the barriers present difficulties for instantiating a unified moral, relational framework in practice for data-driven algorithmic systems. We conclude by discussing ways of weakening the barriers in order to do so.

CCS CONCEPTS

• **Social and professional topics** → *Government technology policy*; **Codes of ethics**; **Socio-technical systems**.

KEYWORDS

accountability, relationality, moral philosophy, robustness, data-driven algorithmic systems

ACM Reference Format:
A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533150

*Equal contribution

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533150

1 INTRODUCTION

In 1996, Nissenbaum [95] warned of the erosion of accountability due to four barriers inimical to societies increasingly reliant on computerized systems. These barriers are: *many hands*, to refer to the problem of attributing moral responsibility for outcomes caused by multiple moral actors; "*bugs*," the way software developers might shrug off responsibility by suggesting software errors are unavoidable; *computer as scapegoat*, the shifting of blame to computers as if they were moral actors; and *ownership without liability*, the free pass to the software industry to deny responsibility, particularly via shrink-wrap and click-wrap Terms of Service agreements. Today, twenty-five years later, significant work has been done to address the four barriers through developments in professional practices of computer science [61, 127], organizational management [62], and civil law [40, 91]; however, the effort to restore accountability remains incomplete. In the interim, the nature of computerized systems has been radically transformed by the ascendance of data-driven algorithmic systems[1]—e.g., machine learning (ML) and artificial intelligence (AI)—which either have replaced or complemented rule-based software systems, or have been incorporated within them as essential elements [12, 26, 75, 92, 141].

The resurgent interest in accountability is therefore timely for a world in which data-driven algorithmic systems are ubiquitous.[2] In domains as varied as finance, criminal justice, medicine, advertising, hiring, manufacturing, and agriculture, these systems are simultaneously treated as revolutionary, adopted in high-stakes decision software and machines [5–7, 69, 89], and as novelties [68]. The failure to comprehensively establish accountability within computational systems through the 1990s and 2000s has thus left contemporary societies just as vulnerable to the dissipation of accountability, with even more at stake. We remain in need of conceptual, technical, and institutional mechanisms to assess how to achieve accountability for the harmful consequences of data-driven algorithmic systems—mechanisms that address both *whom* to hold accountable and *how* to hold them accountable for the legally cognizable harms of injury, property loss, and workplace hazards, and the not-yet-legally-cognizable harms increasingly associated with these systems, such as privacy violations [28], manipulative practices [4, 73], and automation-driven discrimination [5].

[1] Since rule-based software systems are also "algorithmic," we specify which of the meanings we intend in settings where the context does not disambiguate.
[2] We adopt the term "Algorithmic Society" as used in Balkin [9].

In light of growing concerns over accountability in computing, our paper revisits Nissenbaum's "four barriers to accountability" to assess whether insights from that work remain relevant to data-driven algorithmic systems, and to consider how the ascendance of such systems complicates, challenges, and demands more of sociotechnical, philosophical, and regulatory work. We first provide context on recent developments in standards of care, law and policy, and computer science that are necessary for our analysis (Section 1.1). Equipped with this background, we recapitulate the elements of moral philosophy on which Nissenbaum [95] depended (Section 2.1), and discuss how this moral conception of accountability can be unified with Bovens's relational definition of accountability in political theory [20], which has drawn recent attention in AI ethics scholarship. In particular, we contend that moral and relational accountability can be brought together to illuminate the necessary parameters of an accountability framework for data-driven algorithmic systems—determining *who* is accountable, *for what*, *to whom*, and *under which circumstances* (Section 2.2). To instantiate such a framework, however, requires recognizing the ways in which data-driven algorithmic systems specifically make determining these parameters challenging. We therefore update Nissenbaum's four barriers to accountability in relation to these systems, and clarify the ways that each barrier obscures and complicates realizing a moral, relational accountability framework in practice (Section 3). Finally, we conclude by suggesting ways of weakening these barriers to accountability, thereby strengthening accountability practices for the entire field (Section 4).

1.1 Contemporary Interventions in Technological Accountability

Re-visiting the four barriers requires engaging with the significant body of work on accountability produced in the interim. Rather than comprehensively reviewing existing literature—an undertaking already addressed in, e.g., Wieringa [137] and Kohli et al. [71]—we highlight three areas of work that we find useful for our analysis:

Standards of care. These play a crucial role in building a culture of accountability— establishing best practices and formal guidelines for ensuring that concrete practices align with agreed-upon values (e.g., safety). In engineering, standards of care dictate the behaviors and outputs expected of sound work. For data-driven algorithmic systems in particular, they have taken the form of annotations [13], audits [5], and frameworks concerning the appropriate use of data and other artifacts, which are often developed and used in the production of AI/ML systems [22, 52, 60, 84, 88, 112]. Taken together, these standards of care support accountability by making the intentions and expectations around such systems concrete; they provide a baseline against which one can evaluate deviations from expected behavior and, accordingly, are used to review and contest the legitimacy of specific applications of data-driven techniques. Some scholars have re-framed such standards around harmed and vulnerable parties [87, 103]. This work makes clear that standards of care, while important for developing actionable notions of accountability, do not guarantee accountability on their own [37, 128]. Algorithmic impact assessments attempt to fill this gap [90]. They task practitioners with assessing new technologies in terms of their anticipated impacts [87, 108], and formalize accountability relationships in ways that may systematically address and correct algorithmic harms.

Law and policy. Literature on data-driven algorithmic systems generally concerns AI/ML-related harms and corresponding interventions. Work on liability spans both anticipated harms related to new or forthcoming data-driven technology, including autonomous vehicles and robotics [2, 6, 31, 39, 121], and not-yet-legally-cognizable harms, such as unfair discrimination due to demographically-imbalanced, biased, or otherwise-discredited training data [57, 96, 133], privacy violations [28, 34, 65], and manipulation [73]. Regulatory and administrative scholarship tends to analyze data-driven algorithmic systems in relation to legislation and policy that predates many AI/ML technological developments [35, 92, 105, 110, 130, 136]. That said, recent regulatory interventions, including GDPR (the nascent, yet wide-reaching data-privacy policy in the EU [54, 66, 132]) and the California Consumer Privacy Act of 2018 [11], which have also been applied to AI/ML systems, are increasingly represented within the law and policy literature.

Law and policy approaches tend to focus on transparency, which is of broad import in democratic governance and is intimately connected to accountability [91]. Transparency is necessary for identifying responsible parties (in order to attribute harms to those who are responsible for them), and necessary for identifying the sources of these harms and potential mitigations [37]. Work in this area spans a range of urgent concerns surrounding lack of transparency in data-driven algorithmic systems. These include the obfuscation of data provenance [80, 131], particularly caused by the concentration of data ownership within data brokers [41, 76, 140], and insufficient transparency of algorithms and models, which contributes to the inscrutability of automated decisions [29, 75, 77]. Critics have argued that outsourcing legal decisions to automated tools, particularly data-driven tools that obscure underlying decision logic, can create a crisis of legitimacy in democratic decision-making [25, 27, 92, 126].

Computer science. Research in AI/ML has increasingly treated accountability as a topic for scholarly inquiry. In updating Nissenbaum's barriers, we address cases in which researchers explicitly recognize the relationship between their work and accountability [67]—namely, in auditing and transparency—and work on *robustness*, which we identify as having significant implications for accountability, even when this work itself does not explicitly make the connection. Recent work on audits underscores the importance of being able to analyze algorithmic outputs to detect and correct for the harm of unfair discrimination [3, 103]. Transparency tends to be treated as a property of models, particularly whether a model is interpretable or explainable to relevant stakeholders [14, 38, 48]. More recently, computational work has begun to take a more expansive view of transparency, applying it to other parts of the ML pipeline, such as problem formulation, data provenance, and model selection choices [33, 47, 74, 114, 115].

Lastly, often overlooked, *robustness* draws attention to whether a model behaves as expected under likely, unlikely, anomalous, or adversarial conditions. Designing for and evaluating robustness implicates accountability, as it requires researchers to define their expectations of model performance rigorously; this in turn

encourages inquiry into how to prevent deviations from those expectations, and to identify (and ideally correct for) such deviations. Robustness thus encompasses work in AI/ML that aims to achieve theoretical guarantees in practice [85, 139, 142], and work that, even in the absence of such guarantees, produces models with reproducible empirical behavior [19, 102]. Robustness also includes the ability for models to generalize beyond the data on which they were trained [59, 94], ranging from natural cases of distribution shift [70, 97] to handling the presence of adversaries that are trying to game model outputs [53, 98, 122].

2 CONCEPTUAL FRAMING

The conceptual framing of accountability for this paper draws from two sources of scholarship: 1) moral philosophy, which construes accountability as a relationship between and among multiple actors; and 2) political theory and the social sciences, largely focusing on work by Mark Bovens, whose framework for identifying accountability relationships has been particularly influential in contemporary scholarship on "algorithmic accountability"[3] [63, 74, 137].

2.1 Accountability in Moral Philosophy

Numerous efforts in moral philosophy have sought to develop a rigorous conception of accountability. We focus on two threads in the literature, *blameworthiness* and *relationships between moral actors*, and correspondences between the two.

Blame. Nissenbaum [95] anticipated a problem of diminishing accountability as societies become increasingly dependent on computerized systems. She attributed this likelihood to the emergence of barriers to accountability in computerized society, and turned to philosopher Joel Feinberg's work to explain how and why these barriers are prone to arise: Blame, defined in terms of *causation* and *faultiness*, is assigned to moral agents for harms they have caused due to faulty actions [43, 44].[4] Following Feinberg, Nissenbaum conceives of actors as accountable when they step forward to answer for harms for which they are blameworthy. Her concern was that in computerized societies too many circumstances would arise where no one would step forward to acknowledge blame for harm, whether due to genuine puzzlement or intentional avoidance. Accordingly, the barriers to accountability that she identifies arise because the conditions of accountability are systematically obscured, due, at some times, to circumstances surrounding computerization and, at other times, to a societal breakdown in confronting willful failures. *Many hands* obscures lines of causal responsibility (Section 3.1); *"bugs"* obscures the classification of errors as instances of faulty action (Section 3.2) ; *scapegoating computers* obscures answerable moral actors by misleadingly or mistakenly attributing moral agency to non-moral causes (Section 3.3); and *ownership without liability* bluntly severs accountability from blame (Section 3.4) .

[3]We discuss concerns with this phrase in Section 3.3 (*scapegoat*).

[4]Neither of these elements is straightforward—in fact, the subjects of centuries of philosophical and legal thinking. Faultiness, e.g., presumes free agency—a concept whose metaphysical character and role in moral attribution has been the subject of long debate—and is a basic concept in all legal systems that informs judgements of legal liability (categorizing harmful actions as intentional, reckless, and negligent) [42, 43].

Relationality. An alternative conception of accountability expands the focus to consider responsibility in light of the relationships between moral actors. Watson [135], for example, argues that responsibility should cover more than *attributablility*, a property assigned to an actor for bringing about a given outcome [123]. A second dimension, which he calls *accountability*, situates responsibility in a *relationship among actors*. For Watson, "Holding people responsible is not just a matter of the relation of an individual to her behavior; it also involves a social setting in which we demand (require) certain conduct from one another and respond adversely to another's failures to comply with these demands" [135, p 229]. Other work, including T.M. Scanlon's theory of responsibility, provides accounts of both *being responsible* and *being held responsible*, where the latter describes situations when parties violate relationship-defined norms [106, 113]. Accordingly, the characteristics of a harmed party might dictate whether, or what, accountability is needed. For instance, if one causes harm in self defense, there may be no moral imperative to hold them accountable.

This work attempts to situate accountability in the social, political, institutional, and interpersonal relationships in which we are enmeshed. Accordingly, the relationship-defined obligations we have to one another—as spouses, citizens, employees, friends, etc.—may dictate what it is we are responsible for, as well as the types and degrees of accountability we can expect. By situating accountability not just as *attributability* between action and actor, but instead within a social framework, some of what has come out of the so-called "narrow" notion of accountability in political theory (discussed below in Section 2.2) can be derived from the vantage of a more "pure" moral philosophy. Rather than formally pursuing this derivation here, we instead simply suggest that these notions of accountability need not be framed as alternatives to one another. Moral philosophy offers concepts through which a given relational framing—be it interpersonal, institutional, or political—can be said to be legitimate and ethically viable. Similarly, for practitioners holding a variety of organizational positions (in relation to one another), the moral responsibilities that individuals hold can shape the ethical obligations and specific forms of accountability at play.

2.2 Accountability in Political Theory and the Social Sciences

The work in moral philosophy discussed above aligns with work on accountability as a property of social structures [51], which holds it to be relational—not merely as a requirement on an accountable party to "own up" to blameworthy action as an obligation *to* another. In the past few years, "algorithmic accountability" has attracted growing interest in approaches that are institutional or structural in character. The work of political scientist Mark Bovens, particularly what he has labeled, a "narrow definition" [20, 21], has informed recent literature on accountability for "algorithmic systems" [137]. Prompted by a concern that newly formed governmental structures and public authorities in the European Union lack "appropriate accountability regimes" [21, p. 447], Bovens proposed that accountability obtains between two key roles: an *accountable actor* and a *forum*. Under certain conditions, or in the wake of certain incidents, accountability exists when an accountable actor has an enforceable obligation to a forum to explain and justify itself—to address

a forum's questions and judgments and possibly suffer sanctions. Bovens calls this a "relational" definition because it locates accountability in a social relation between those occupying one role (e.g. governmental department, a public authority, or a person acting in an official capacity) and another (e.g., a different governmental entity, oversight committee, or even an individual acting in a relevant capacity, e.g. journalist). We read Bovens as gesturing toward four key parameters in any relational accountability framework for which appropriate values need to be specified:

Who is accountable?: *Accountable actors* may include those who are not directly responsible for harm (e.g., engineers) but are designated as accountable (or liable) because of their deep pockets, capacities to render explanations, or positions in organizational hierarchies, such as corporate officers or government procurers of data-driven systems.

For what?: Beyond legally-cognizable harms (e.g., bodily injury, property damage), harms particularly associated with data-driven algorithmic systems include privacy violations [28], automation-driven unfair discrimination [5], autonomy losses due to manipulation [4, 73], and any number of emergent harms associated with novel technologies and their deployment.

To whom?: The members of the *forum* may not just include those who are themselves harmed (or placed in harm's way through heightened risk). They may also include those deputized to represent and advocate on behalf of vulnerable parties, such as lawyers and public or special interest advocacy groups. Beyond direct advocates, these may include groups and individuals in oversight capacities such as journalists, elected officials, government agencies, professional societies, or the many publics which coalesce around particular matters of concern [86].

Under which circumstances?: This concerns the nature of the obligation—what *accountable actors* may owe to the forum (to explain, be judged, and address questions and challenges). For example, Moss et al. [90] describes an array of components that constitute accountability within impact assessment frameworks, noting that the specific obligations an actor owes to a forum depend on the norms of that relationship.

Bringing together the moral and the relational. Proponents of Bovens's relational framework claim that it illuminates the sociopolitical stakes of transparency and explainability, showing why these concepts are necessary for any accountability framework for algorithmic societies, even though they are ultimately not sufficient to constitute accountability in and of themselves [137]. Moreover, by defining actors' roles and capacities in terms of the respective sociopolitical structures in which we live, Bovens's framework is *not* directed at the rights and obligations we have to one another as bare moral actors. We note that bringing together Bovens's relational definition with the moral conception of accountability can help clarify the scope of possible values for the framework's parameters: Those who have caused or contributed to harm through faulty action are contenders for the class of *accountable actors*, and those who have suffered harm (and/or their representatives) deserve a place among the members of the *forum*. This point shows a confluence between accountability as answerability for blameworthy action, and accountability as a social arrangement. Being blameworthy for harm is (almost always) a sufficient condition for being designated an accountable actor; being harmed through blameworthy action is (almost always) a sufficient condition for being designated a member of the forum, empowered to demand explanations. These two conceptions do not stand against one another as alternative solutions to the same problem; they are solutions to different problems that intersect in constructive ways.

Nevertheless, hard work remains to explain and justify concrete, appropriate values for these parameters, and to construct pervasive structures for accountability through context-bound contestation [87]. In Section 3 below, we demonstrate how data-driven algorithmic systems heighten the barriers to accountability by further obscuring conditions of responsibility and fault, which in turn presents challenges for instantiating the four parameters of a moral, relational accountability framework.

3 REVISITING THE FOUR BARRIERS TO ACCOUNTABILITY

In a typical scenario in which software is integrated into a functional system—fully or partially displacing groups of human actors—accountability could be displaced along with human actors who are its bearers. The cumulative effect of such displacements is the increasing incidence of harmful outcomes for which no one answers, whether these outcomes are major or minor, immediate or long-term, or accrue to individuals or to societies. Resuscitating accountability is no simple task because computerization sets up particularly troublesome barriers to accountability: *Many hands* (3.1), *"Bugs"* (3.2), *The computer as scapegoat* (3.3), and *Ownership without liability* (3.4) [95]. These interdependent barriers are not necessarily an essential quality of computer software. Rather, they are a consequence of how software is produced, integrated into institutions, and embedded within physical systems; they are a function of the wonderment and mystique that has grown around computerization, and the prevailing political economy within which the computer and information industries have thrived. In the sections that follow, we revisit the barriers with an eye turned toward their implications amidst the massive growth and adoption of data-driven algorithmic technologies. We provide examples of the barriers in action and defer discussion of how the barriers can be weakened to Section 4.

3.1 The Problem of *Many Hands*

The barrier of *many hands* arises due to the large number of actors often involved in the design, development, and deployment of complex computerized systems. When such systems cause harm, it may be difficult to isolate the component(s) at its source and the agents responsible: "Where a mishap is the work of 'many hands,' it may not be obvious who is to blame because frequently its most salient and immediate causal antecedents do not converge with its locus of decision making" [95, p. 29]. Nissenbaum further analyzes the difficulty of *many hands* by showing how it operates at four different levels: 1) software is produced in institutional, often corporate, settings in which there is no actor responsible for all development decisions; 2) within these settings, multiple, diffuse groups of engineers contribute to different segments or modules of the overall deployed system, which additionally often depends on software implemented by other actors (in today's landscape, this may result in licensed or freely-available open-source software); 3)

individual software systems often interact with or depend on other software systems, which themselves may be unreliable or present interoperability issues; 4) hardware, not just software, often contributes to overall system function, particularly in cyber-physical systems, and it can be difficult to pinpoint if harms occur due to issues with the code, the physical machine, or the interface between the two. Any and all of these four levels of *many hands* problems can operate simultaneously, further obscuring the source of blame.

These difficulties at the heart of the *many hands* problem persist, further complicated in numerous ways now that computer systems are ubiquitous rather than merely ascendant. We focus on how data-driven algorithmic systems complicate this barrier with novel challenges using two illustrative examples: 1) The *ML pipeline*—the multi-stage process by which machine-learned models are designed, trained, evaluated, deployed, and monitored; 2) Reliance of contemporary data-driven algorithmic systems on the *composability of openly-available ML toolkits and benchmarking suites*; these toolkits, often developed and maintained by large tech companies, tend to be advertised as general- or multi-purpose, and are frequently (mis)used in specific, narrow applications.

The ML pipeline. The ML pipeline is a dynamic series of steps, each of which can involve multiple groups of actors, including designers, engineers, managers, researchers, and data scientists. The pipeline typically starts with problem formulation and, in commercial settings, results in the deployment and continued monitoring of a trained model [99]. Problem formulation involves the collection, selection, or curation of a dataset, followed by the operationalization of a concrete task to learn, such as classifying loan-granting decisions or generating natural-language text. The actors responsible for formulation may hand off their work to others responsible for implementation—choosing the type of model and the learning procedure to use for model training. In selecting the type of model, these actors may custom-design their own architecture, or may defer to a pre-existing one, such as an off-the-shelf neural network, which has been designed by others, possibly at another company or institution. Thereafter, training and evaluation begin, in which a group of developers run training many times, perhaps with multiple combinations of model types, training procedures, and hyperparameter values. These developers compare trained models, from which they select some "best"-performing model (or ensemble of models), where "best" is informed by a quantitative metric they have adopted, such as mean overall test accuracy. These stages, from formulation to evaluation, are often repeated dynamically: Until the model passes the threshold of developer-specified performance criteria, the process can cycle from re-modeling to tuning. If the model is deployed in practice, there is yet another set of actors who monitor the model's ongoing behavior, ensuring that its behavior aligns with expectations developed during training and evaluation.

Each stage of the ML pipeline involves numerous actors—in fact, potentially *indefinitely* many actors if the pipeline employs third-party model architectures or ML toolkits, which we discuss below.[5] Thus, in practice, if a trained model causes harms, it can be extremely challenging to tease out particular actors who should answer for them. For example, harms could originate from how actors operationalize the learning task at the start of the pipeline [30], move from high-level abstraction to concrete implementation [109], or select hyperparameters or random seeds during model selection [33, 47, 114]. Blame could lie with actors in any part of the pipeline, or some combination thereof whose faulty actions may have been causally responsible for harm. Bias, for example, could creep in early, from the choice of dataset, and accumulate and become magnified further downstream during model selection. In other words, the diffuse and dynamic nature of the pipeline makes locating accountability extremely challenging. This can be understood as an issue of transparency—beyond the specific the problem of model interpretability—concerning *who* is responsible *for what*, and *how* this can be related to overarching accountability with respect to a model's ultimate use in practice [74].[6]

Multi-purpose toolkits. Practitioners and researchers often do not code model architectures or optimization algorithms from scratch. Just as Nissenbaum highlighted the integration of third-party software modules as the indefinite expansion of *many hands*, we note here that builders of data-driven algorithmic systems often rely on toolkits produced by others. To decrease the amount of time and money spent iterating the ML pipeline, these actors depend on the investment of tech companies with vast resources and large, concentrated pools of technical talent to develop and release efficient, correct, comprehensive, and user-friendly libraries of algorithm implementations, model architectures, and benchmark datasets [1, 83, 100]. Unlike more traditional modules, which only tend to contain reusable software algorithms, ML toolkits often also include large-scale, pre-trained models. Large companies train and release such models, like BERT [36], which smaller companies and individuals can use out-of-the-box or fine-tune for particular use cases. Since these pre-trained models are often intended for downstream use by users different from their developers, they are designed for a multiplicity of applications (i.e., to be general-purpose). However, users employ pre-trained models in specific domains; there is a gap between general design goals and specific deployment intentions, which has been shown can bring about bias-related harms. Determining blame for these types of harms is far from simple. For example, if intended use is under-specified, blame could lie at least partially with the pre-trained model's creator. Compounding this problem is the fact that ML presents a recursive turn in the *many hands* problem Nissenbaum highlighted, in that many ML systems incorporate pre-trained components that are, themselves, the product of *many hands*. Nevertheless, tracing such harms presents an addressable technical challenge, not an insurmountable epistemological barrier.

In relation to a moral, relational accountability framework, this barrier obscures …

Who is accountable: *Many hands* is central to identifying an accountable actor within Bovens's framework [20]. This problem has long characterized challenges in holding corporate actors, institutions, and organizations accountable, and while it certainly constituted a barrier to accountability in 1996 [95], it has only become

[5] Participatory design further expands the set of *many hands* to end-user stakeholders [115], illustrating an additional manifestation of the barrier: when harms occur, it is possible to shift blame to harmed end-users who were involved in the ML pipeline.

[6] This indicates why transparency in the form of model intepretability may be important, but is ultimately not sufficient, for identifying actors accountable for harms.

more difficult to understand who is accountable in an algorithmic society. Code reuse—taken as a virtue in software development—has now been extended to model reuse, in turn generating a host of problems for equity and reliability by making it difficult to identify all the actors who contributed to components of an ML pipeline. Knowing who is responsible for these components as they are repurposed, as well as who ought to be responsible for incorporating those components into a downstream system, becomes prohibitively difficult for a forum to ascertain on its own, let alone for it to demand any explanations or changes in actors' behavior.

For what: The problem of *many hands* extends the above question to determining *what* an actor might be accountable for in relation to harms, in that it is hard to isolate which part of an ML pipeline actually contributes to an error or harm. Repurposed models may introduce dataset imbalances and proxies for protected categories without adequate scrutiny (or even the opportunity for scrutiny) by those assembling downstream components of a system. This raises questions of appropriate use, wherein it is difficult to tease apart the responsibility of those who produced a component to adequately stipulate the limits of its appropriate use and the responsibility of those who use that component to ensure it is appropriate for the uses to which they are putting it.

To whom: *Many hands* is primarily a barrier to knowing *who* is accountable, but it is also a barrier to knowing *to whom* those accountable actors are accountable where, for example, a differential error rate may exist for some population $\mathcal{P}$, but a specific harm occurs for an individual $p \in \mathcal{P}$. In such a case, it is difficult to determine whether accountability ought to be rendered to $\mathcal{P}$, because of the heightened risk of harm to which the entire population has been exposed, or only to p, who suffered harm because of their membership in $\mathcal{P}$. This is a *many hands* problem because of the difficulty in knowing where within the ML pipeline risk was produced for the group, e.g. through training or dataset imbalances, and where it was produced for individuals, e.g. through implementation choices.

Under which circumstances: The problem of *Many hands* presents a barrier even when standards of care exist, as it is difficult for actors to know precisely whom they should exercise that care toward (see "to whom" above). Standards of care, which are grounded in normative assumptions about appropriate component (re)use, are less straightforward to develop where *many hands* are involved, as social practices which link actors together are obscured throughout the ML pipeline.

3.2 *"Bugs"*

Nissenbaum uses the term *"bug"* to cover a variety of issues common to software, including "modeling, design, and coding errors." *"Bugs"* are said to be "inevitable," "pervasive," and "endemic to programming," "natural hazards of any substantial system" [95, p. 32]. Even with software debuggers and verification tools that can assure correctness, *"bugs"* emerge and cause unpredictable behavior when software systems are deployed and integrated with each other in the real world [82, 117]. The rhetorical power of *"bugs"* is that they are predictable in their unpredictability; they serve as a barrier to accountability because they cannot be helped (except in obvious cases), and therefore are often treated as an accepted "consequence of a glorious technology for which we hold no one accountable" [95, p. 34]. What we consider to be the "inevitable" can change over time as technology evolves, with certain types of *"bugs"* spilling over into the avoidable. For example, evolving norms and new debugging tools can rebrand the "inevitable" to be sloppy or negligent implementation, at which point programmers can be held to account for such errors. Similarly, the advent of data-driven algorithmic systems has indicated that this malleability also extends in the other direction: New technological capabilities can both contract and expand what we consider "inevitable" *"buggy"* behavior. That is, while these systems contain *"bugs"* of the "modeling, design, and coding" varieties that Nissenbaum describes for rule-based programs, the statistical nature of data-driven systems presents additional types of harm-inducing errors, which may present an additional barrier to accountability.[7] Where misclassifications, statistical error, and nondeterministic outputs cause harm—and are presented as inevitable and unavoidable—may impede the attribution of blame.

In 1996, it may have been evident that labeling certain errors as *"bugs"* was a mere ploy to dodge blame. Today, certain types of errors are more plausibly asserted to be an inherent part of ML, attributable to its statistical nature. Misclassification, statistical error, and nondeterminism seem to turn the notion of *"bug"* on its head: Indeed, many experts would as readily call these *features* of machine learning, not *"bugs"*.[8] Nevertheless, regardless of where one attempts to draw the line, these errors share common elements with the *"bugs"* Nissenbaum describes—namely, they undermine our ability to reason, conclusively, about causality and fault. Insofar as they are accepted as an "inevitable," "pervasive," and "consequence of a glorious technology," they constitute a barrier to accountability [95]. Below, we illustrate this point with concrete instances of *"bugs"* deemed unavoidable in data-driven algorithmic systems.

Faulty modeling premises. As discussed in Section 3.1, data-driven algorithmic systems require significant modeling decisions prior to implementation. For example, choosing a model to learn necessarily involves abstraction and can have significant ramifications [99, 109]. Assumptions during this stage of the ML pipeline can bias the resulting computational solution space in a particular direction [49], for example, assuming a linear model is sufficient to capture patterns in data precludes the possibility of modeling non-linearities. When such biases involve over-simplified or faulty reasoning, they can result in model mis-specification and the introduction of "modeling error bugs." Such mis-specifications may include the assumption that values like fairness and accuracy are correctly modeled as a trade-off to be optimized [30], and that physical characteristics can serve as legitimate classification signals for identifying criminals [138] or inferring sexual orientation [119, 134]. More generally, a common modeling error may arise from assuming, in the first place, that a problem is amenable to classification—that it is possible to divide data examples into separable categories [116, 120]. Even if it is possible to train mis-specified models like these to behave "accurately" (i.e., to return better-than-chance results after learning these tasks), conclusions drawn from false premises will be unsound [30]. If modeling assumptions are unclear or elided, an actor may evade accountability by blaming

[7] Of course, statistical software is not new to ML; however, the proliferation of data-driven algorithmic systems has clarified the prevalence of such errors.
[8] We return to this in Section 3.3 (*scapegoat*) and is why we leave *"bugs"* in quotes.

inexplicable, unavoidable *"bugs"* endemic to computer software instead of taking responsibility for otherwise opaque errors.

Individual errors. Even if one's premises are not faulty, the ML pipeline can still produce models that cause harm. Trained ML models exhibit errors that can harm individuals if their effects, for example, violate privacy or cause manipulation [45, 73, 81, 93]. ML has several techniques to quantify and minimize error [17, 55], and yet even the most robust, well-trained models report imperfect accuracy. In fact, a model that achieves 100% accuracy is usually considered suspect, likely over-fit to the training data and to exhibit poor performance when presented with new examples [56, 104, 118]. Therefore, when individual errors occur, they can be treated as inevitable, just like the *"bugs"* Nissenbaum describes, displacing responsibility for the harms such errors cause affected individuals.

Bad model performance. Unexpectedly bad overall model performance can likewise be excused as a *"bug,"* rather than a blameworthy error. Consider a hypothetical example of a (well-formulated) computer vision system used to detect skin cancer, whose training and evaluation indicate will have an accuracy rate of 94%. Once deployed, if the model coheres with (or even out-performs) its promised performance, then developers can claim that any misclassifications were expected.[9] Since expected accuracy is a probabilistic claim about what is likely to occur, deviations from expectation can and do occur. When monitoring a deployed model, over time, if this deviation yields a substantial decrease in expected accuracy, developers may dodge accountability by ascribing the failure to the amorphous category of *"bug"*, instead of admitting that it resulted from human negligence, poor generalization, distribution shift, or other faulty behavior.

In relation to a moral, relational accountability framework, this barrier obscures …

Who is accountable: Accountability for *"bugs,"* even within the expanded definition of *"bugs"* provided above, emerges from specific regulatory regimes, corporate compliance practices, and contracting relationships. Civil law has a crucial role in determining the relationship between forums and responsibilized actors, which is often inflected by those who have the capacity to intervene or have benefitted from a particular action. *"Bugs"* present a particular challenge to determining who is accountable when they are seen as endemic to ML, or as produced by non-determinism inherent to the domain in which an algorithmic system is deployed (a challenge shared by the *scapegoating* barrier).

For what: *"Bugs"* remain a barrier because of the difficulty they pose to actors and forums trying to specify whether individual errors, bad model performance, faulty assumptions, or other mistakes contributed to a harm.

To whom: *"Bugs"* may affect an entire class of individuals, a community, or all of society, but evidence of harm may only accrue at the level of a specific individual, presenting a barrier for actors and forums interested in knowing to whom accountability ought to be rendered.

Under which circumstances: Algorithmic systems inevitably rely on some degree of abstraction and make specific assumptions about the underlying nature of the phenomena they model [30, 109]. Under circumstances of imperfect information about every possible aspect of a data-driven algorithmic system (which is most of the circumstances outside the lab), *"bugs"* of the character described above may exist and contribute to this barrier to accountability.

3.3 *The Computer as Scapegoat*

Blaming a computer may pose a barrier to accountability, because "having found one explanation for an error or injury, the further role and responsibility of human agents tend to be underestimated" [95, p. 34][39]. To explain why people could plausibly blame computers for wrongdoing, Nissenbaum cites the role computers may play in "tasks previously performed by humans in positions of responsibility;" whereas before the human would be indicated as the blameworthy party, the computer has now taken up that role. And yet, even as computer systems have become immediate causal antecedents to an increasing number of harms, they lack moral agency and thus cannot be the bearers of moral blame [95]. In this section, we discuss how *scapegoating the computer* has become even more complicated in the landscape of ubiquitous data-driven algorithmic systems. In the examples below, the system is made to bear the sins of the responsible party, the individual or the institution that has agency and is capable of carrying moral blame.

Moral agency. As data-driven algorithmic systems have become pervasive in life-critical contexts, there has been a corresponding tendency to anthropomorphize and view technological processes as akin to human cognition [15, 101, 125]. These systems are described by their developers and commentators as intelligent, implying that they have agency as autonomous actors and thus rhetorically positioning them as blameworthy for error. However, directing blame toward data-driven algorithmic systems effectively imbues them with moral agency, ascribing them the ability to act intentionally [107]. Nissenbaum likens blaming a computer to blaming a bullet in a shooting: While the bullet can be said to play an active, causal role, it cannot be said to have been intentional in its behavior. In the same vein, a data-driven algorithmic system may play a central role in life-critical decisions, and may even be said to *make a choice* in a particular task, but a choice lacking deliberate intention, a precondition for moral agency [107].[10]

"Accountable algorithms". This popular banner-phrase makes *algorithms* the subject of accountability [75], even though algorithms are not bearers of moral agency and, by extension, moral responsibility. It places responsibility on technology, not its developers, owners or operators, and it reduces accountability to a piecemeal, procedural quality that can be inferred from technology, rather than a normative concept that has to do with the moral obligations that people have toward one another. The phrase further occludes proper attribution of accountability by fixating attention on algorithms rather than on systems that are deployed in practice, within and through which algorithms function [32]. When, for example, studies of fairness in AI/ML-assisted judicial bail decisions fixate

[9] Individual errors can pose additional challenges for accountability: The model may still overall exhibit an expected degree of error (i.e., be within a margin of error), for which it is possible to scapegoat the statistical nature of ML (Section 3.3).

[10] This is consistent with scholarship in legal theory concerning AI, algorithms, agency, and personhood [8, 15, 23, 24, 58, 78, 129].

on respective algorithms, they fail to capture key inequities that are systemic in complex sociotechnical systems, of which AI/ML techniques are just one part [10].

Mathematical guarantees. Directing blame away from people and corporations can be either strategic or inadvertent. In some cases, a group of harmed individuals does not know whom to blame (*many hands*) and settles on blaming the system. In others, scapegoating the system can be a way by which a moral actor dodges and dissipates public ire, for example, in the now-canonical example of Northpointe exhibiting bias in its risk-assessment tool [7]. Rather than attributing this bias to a mistake or *"bug,"* Northpointe blamed the fundamental incompatibility of different algorithmic operationalizations of fairness as the source of the problem (and pointed to a specific measure, for which bias was not detectable, as evidence of blamelessness). Reliance on mathematical guarantees can reinforce barriers to accountability and divert attention away from its appropriate subjects. One can see this when a given system has a theoretically-guaranteed (and empirically-verified) upper bound on its error. If the system behaves within its guaranteed margin of error, it becomes possible to treat that margin as an immutable attribute of the system (rather than, more appropriately, the result of human-made decisions), and to scapegoat the system for any particular errors that fall within this margin.

Let us consider the same case we discussed for the problem of individual errors in *"Bugs:"* The engineers show that a system is 94% accurate for tumor detection, and validate that this is in fact the case in practice. Above, we talked about this example in terms of individual errors, for which responsibility for harm could be excused due to *"buggy"* behavior. Rather than analyzing behavior at this level of individual decisions, one can also examine the behavior *of the model overall*. If the frequency of mis-classifications is within the model's guaranteed error rate, the engineers could attempt to excuse all resulting harms by gesturing to the fact that the model is performing *exactly as expected*. In short, satisfying mathematical guarantees can serve as a *scapegoat* because pointing to mathematical claims satisfied at the model-level can serve to obscure the need to account for harms that occur at the individual-decision level.[11]

Non-determinism. When data-driven algorithmic systems err, their errors can be attributed to the stochastic, non-deterministic components of either the system itself or the phenomena the system is modeling. In particular, systems that involve ML involve randomization, for example, by shuffling the order in which training data examples are presented to an algorithm. While such features of ML algorithms may seem like technical minutiae, in fact, they introduce stochasticity into the outputs of machine-learned models: Training the same model architecture on the same dataset with the same algorithm—but changing the order in which the training data are supplied to the algorithm—can yield models that behave differently in practice. For example, as Forde et al. [47] shows, changing the order that the data examples are presented to train a tumor-detection model can lead to surprisingly variable performance. The relationship between training-data-ordering and resulting variance in model performance is under-explored in the technical literature. Thus, such differences in model performance are often attributed to an inherent stochasticity in ML. The randomization used in ML systems—randomization on which these systems depend—becomes a *scapegoat* for the harms it may cause, such as missed tumor detection. In attributing the harms to mathematical chance, attention is drawn away from appropriate accountable agents.

In relation to a moral, relational accountability framework, this barrier obscures ...

Who is accountable: Similar actors are accountable as those described in *"Bugs,"* although the barrier presented by *scapegoating* is embedded in its implicit suggestion that entities are accountable, rather than those who are the responsible actors (see the nebulousness of *many hands*), or that no responsible actor can be found because a harm occurred through randomness or chance.

For what: *Scapegoating* produces barriers to understanding the *for what* of accountability in identical ways as described above in *"Bugs,"*. It also contributes an additional difficulty when mathematical guarantees are offered that allow for some minimal degree of undesirable behavior in a system, or the system is characterized as non-deterministic in ways that would indemnify otherwise responsibilized actors from accountability for outcomes stemming from such undesirable behaviors.

To whom and *under which circumstances*: Same as in Section 3.2.

3.4 Ownership without Liability

Nissenbaum [95] highlights a dual trend in the computer industry: 1) strengthening property rights and 2) avoiding liability. Behavioral trends that informed these assertions have persisted in the decades since, with lively public debates over the fit of traditional forms of intellectual property (i.e., copyrights, patents, and trade secrets) to digital products such as software, data, databases, and algorithms [29, 50], and subsequent expensive legal struggles among industry titans [16]. Similarly, we have seen explicit denials of liability expressed in shrink-wrap licenses, carried over into so-called "click-wrap" licenses, and Terms of Service disclaimers accompanying websites, web-based services, mobile apps, Internet of Things devices, content moderation decisions, and the like [28, 72, 79, 124].

Before addressing how we see these trends carry forward in the contemporary landscape, we need to qualify our observations. Property and liability are weighty legal concepts with long histories and rich meanings. Narrowing our view to digital technologies, even before Nissenbaum [95], a robust literature had grown over questions of ownership—questions that have persisted through numerous landmark court cases. Liability, too, is a core legal concept that is increasingly an issue in relation to the products and services of digital industries. It lies outside the scope of this paper to attempt meaningful insights into these concepts as they manifest in scholarship, law, and the courts. However, it is useful to observe broad patterns and anticipate the likely actions of stakeholders. For a start it is not difficult to see how the trends toward strong ownership and weak liability reinforce barriers to accountability, and also to understand why industry incumbents might support them: Liability is costly and strong property rights enrich rights holders and empower them against competitors. Four lines of advocacy

[11] One could see-saw back-and-forth between *"bug"* and *scapegoat* to evade accountability. If satisfying guarantees at the overall model-level is rejected as a rationale for an individual harm, one could claim there is a *"bug;"* if calling an individual decision *"buggy"* is rejected, and the model is classifying within its expected error, one could then displace blame by arguing that the model is behaving according to its specification.

on behalf of industry interests are noted below, supplementary to those discussed in Nissenbaum [95]:

- Third-party providers of data-driven algorithmic systems refuse to expose their systems to scrutiny by independent auditors on grounds of trade secrets [29, 50]. As long as experts maintain that transparency is necessary to evaluate the ML pipeline and AI development, strong property rights that block scrutiny are barriers to accountability.
- Manufacturers and owners of cyber-physical systems, such as robots, Internet of Things devices, drones, and autonomous vehicles, evade liability for harms by shifting blame to environmental factors or humans-in-the-loop [78]. In this respect, the barrier of *ownership without liability* for data-driven algorithmic systems suggests a twist on the problem of *scapegoating* (Section 3.3): treating "the human user as scapegoat"—claiming the user has mis-used an AI- or ML-enabled system in order to obscure responsibility for unclear, under-specified, or deliberately misleading user interfaces or expected use, as has happened with Tesla and accidents concerning its (so-called) "AutoPilot" autonomous driving feature [18].
- Almost without question the computer industry, having metamorphosed into the data industry, has assumed ownership over data passing through its servers [41, 76, 96]. We still do not have clear rules of liability for industry actors when their servers, holding unimaginable quantities of data, are breached [111]. Nor do we have sufficient insight into the completeness, quality, or validity of data, or the means to hold anyone liable for its misuse.
- Technology companies hold unprecedented sway over regulation. Twenty-five years ago, the software industry was already a force to be reckoned with and successfully persuaded Congress that imposing legal constraints would stifle innovation—that societal well-being depended on a nascent industry that could not flourish under excessive regulatory and legal burden.

In relation to a moral, relational accountability framework, this barrier obscures ...

Who is accountable: Having already enumerated above the many difficulties these barriers pose for tracing relationships of accountability, they generally pertain to the problem of *ownership without liability*, as well. Additionally, questions of how liability is adjudicated in practice may obscure who is liable, what kind of liability they hold, or what they are liable for, while leaving intact the ways in which the benefits of data-driven algorithmic systems accrue to their developers, designers, and operators.

For what: *Ownership without liability* affects the very contours of what an actor can be found liable for. However, this does not absolve that actor of their moral responsibility or obviate the need for them to be held accountable for the consequences of their actions, the systems they oversee, or from which they benefit.

To whom: *Ownership without liability* is a barrier to accountability for those who may stand as plaintiffs in civil cases and representatives of those affected.

Under which circumstances: *Ownership without liability* is a barrier where those who suffer a harm lack standing in a court of law. This may be because a harm is not cognizable to courts (see, e.g., Metcalf et al. [86]), the harmed party does not constitute a certifiable class, or the nature of the harm is obscured through the ways harms are foisted onto *scapegoats* or dismissed as *"bugs"*.

4 WEAKENING THE BARRIERS

Nissenbaum [95] warned of a waning culture of accountability—harms befalling individuals, groups, even societies, were being cast merely as sufferers' bad luck. In the previous section, we revisited the four barriers in light of data-driven algorithmic systems and found that the framework still provides a useful lens through which to locate sources contributing to the dissipation of accountability. Weakening the barriers would clear the way for more sound attribution of blame, in turn setting up a stronger societal expectation for blameworthy parties to step forward and take account. But we have also argued that accountability in algorithmic societies involves more: Stepping forward is a necessary component of accountability, but it is insufficient (Section 2). Because the barriers we have described may not all be weakened, even with a firm resolve to identify blameworthy parties, we need more than astute attention on a case-by-case basis. To build a lasting culture of accountability, a necessary supplement involves establishing persistent institutional frameworks for identifying accountable parties (i.e., individuals, groups, or organizations) and for calling them to answer. Simultaneously, such frameworks should invest others with the powers to call these parties to account.

Any technical interventions that the research community has already developed—notably, those that we have emphasized concerning transparency, audits, and robustness—would need to be folded into such a framework, and their use justified in these moral and relational terms. For example, any technical definition of transparency is unlikely to satisfy the needs of all those who comprise a forum and who may hold variable or inconsistent ideas about what it might mean for a model to be "interpretable." Technical assertions of robustness say what expectations are, but leave unanswered the question of the conditions under which deviations from expectations ought to be expected or remedied.[12] Relational treatments of these issues, it would seem, require that the obligation be tuned to the various needs of all members of the forum.

Taking each barrier in turn. A moral and relational accountability framework opens the aperture to addressing *many hands* (Section 3.1). In principle, many, if not all, of the *many hands* could be designated as accountable actors. Deliberate consideration of the *many hands* problem is clearly called for by those who develop licensing agreements relying on normative assumptions about appropriate use and reuse within the ML pipeline, and in articulating engineering best practices empirically against theoretical assumptions of robustness. This includes dataset creators, model developers, decision and control systems designers, vendors, and operators of these systems. Developing rigorous standards of care could help mitigate the problems of inappropriate use of pre-trained models and unclear measures of quality control at different stages of the ML pipeline. For example, robust auditing mechanisms at each stage, rather than approaching audit as an end-to-end concern [103], or worse, as a purely *post hoc* endeavor, could help clarify the relationship between stage-specific issues and resulting harms.

[12] Moreover, if assumptions underlying such assertions are voided when moving from theory to deployment, robustness estimates can degrade in practice.

Addressing various harms, depending on how they are contextualized, can implicate either the barrier of *"bugs"* or *scapegoating the computer* (Sections 3.2 & 3.3). For example, we note that the computer science community could have either treated algorithmic harms due to unfair discrimination as a *"bug"* or blamed them on intrinsic aspects of AI/ML—and yet, it did not. Instead, unfairness has more often been ascribed to biased or imbalanced training data [46, 64]—data that exhibits historical biases that are arguably "pervasive" and "unavoidable." This community could have pursued some "tolerable" degree of unfavorable outcomes in the real world (ideally, in consultation with those adversely impacted), and developed ways of ensuring models met that more "tolerable" specification, under specific conditions. This, notably, would still have allowed developers to evade accountability by *scapegoating* inherent properties of the model as instead deserving of blame.

However, instead of treating unfairness as an aspect of accountability, much technical work on algorithmic fairness has attempted to address unfairness harms by developing training algorithms that are robust to biased input data. The field of algorithmic fairness therefore serves as an example that challenges the narrative of the invulnerability of the barriers. The technical community and its interlocutors have demanded more from ML modelers concerning the treatment of unfair discrimination. The community has set expectations concerning the necessity of interventions to root out and correct for unfairness, thereby weakening the barriers of *scapegoating* or being attributed to *"bugs"*. This example could, and we believe should, encourage similar treatment of other issues like robustness and its relationship to privacy violations, or adversarial ML and its relationship to manipulation.

Lastly, *being liable* is related but not identical to being accountable (Section 3.4). The latter is applied to blameworthy parties who step forward to answer, the former to parties who step forward to compensate victims of harm. Often liability is assigned to those who are found to be blameworthy. If lines of accountability are blurred, for example, as a consequence of the barriers we have discussed, harms due to AI/ML and other data-driven algorithmic systems will be viewed as unfortunate accidents; the cost of "bad luck" will settle on victims. Instead, legal systems have developed approaches, such as strict liability, to compensate victims harmed in certain types of incidents even without a demonstration of faulty behavior. Strict liability assigned to actors who are best positioned to prevent harm is sound policy as it motivates these actors to take extraordinary care with their products. Barriers such as *many hands* make the attribution of blame difficult. Strict liability for a range of harms that are produced by *many hands* would shift the "bad luck" from victims to those best positioned to mitigate and prevent such harms.

Eroding the barriers of accountability is a key societal challenge requiring multiple forms of expertise and, with respect to ML especially, the use of these tools needs to be justified. Just as mature political governance requires durable institutions and formal attributions of rights and duties, we have similar needs for the governance of producers, purveyors, and operators of data-driven algorithmic systems. That is, as we have contended throughout this paper, accountability is moral *and* relational. It depends on social, legal, and political structures that provide legitimacy for the checks actors and forums place on each others' behavior; it depends on the way those checks are internalized as professional, personal, legal, and ethical duties that motivate actors' personal responsibility. Multi- and inter-disciplinary research on accountability, fairness, and transparency—given its potential to bring together an array of expertise focused on themes of equity and justice—is uniquely positioned to help develop a moral, relational accountability framework. Such structures provide legitimacy, as well as the professional codes and standards of care, disciplinary norms, and personal mores that tie moral and relational forms of accountability together. The future work of creating these structures, as noted earlier, is no small undertaking, it lies in the sociopolitical contestations, the hard, deliberative work of living within a pluralistic society, by the many constituencies implicated in any particular computational system.

5 CONCLUSION

In this paper we revisited Nissenbaum's "four barriers" to accountability, with attention to the contemporary moment in which data-driven algorithmic systems have become ubiquitous in consequential decision-making contexts. We have drawn on conceptual framing from Nissenbaum's use of the concept of *blameworthiness* and how it can be aligned with, rather than cast in opposition to, Bovens's work on accountability as a *relational property of social structures* [20, 21]. We have demonstrated how data-driven algorithmic systems heighten the barriers to accountability with regard to determining the conditions of blame, and have looked ahead to how one might endeavor to weaken the barriers. In particular, we have put forward the conditions necessary to satisfy a moral and relational accountability framework, discussed how the development of such a framework would weaken the barriers, and argued that an interdisciplinary and multidisciplinary research community is uniquely positioned to construct such a framework and to develop lines of inquiry to erode the barriers to accountability.

Given our tender historical moment, addressing why these or those parties belong in the forum or in the set of accountable actors, why those obligations are justified, and, of course, evaluating the numerous permutations the relational nature of the approach demands is the provenance of future work. No easy formulations make sense until we have developed a rigorous approach to justification. In our view, this calls for expertise in relevant technologies, moral philosophy, the prevailing political economy of data and computing industries, organizational sociology, current political and regulatory contexts, domain area expertise, and more. It is not that all these are needed all the time; but any of them may be called in to develop linkages between proposed values and social welfare.

ACKNOWLEDGMENTS

The authors would like to thank the Digital Life Initiative at Cornell Tech for its generous support. A. Feder Cooper is additionally supported by the Artificial Intelligence Policy and Practice initiative at Cornell University and the John D. and Catherine T. MacArthur Foundation. Benjamin Laufer is additionally supported by NSF Grant CNS-1704527. Helen Nissenbaum is additionally supported by NSF Grants CNS-1704527 and CNS-1801307, and ICSI Grant H98230-18-D-006.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Kenneth S. Abraham and Robert L. Rabin. 2019. Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era. *Virginia Law Review* 105 (2019), 127–171. Issue 127.

[3] Philip Adler, Casey Falk, Sorelle A. Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54 (2018), 95–122.

[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Long Beach, CA, 8.

[5] Ifeoma Ajunwa. 2021. An Auditing Imperative for Automated Hiring. *Harv. J.L. & Tech.* 34 (2021), 81 pages. Issue 1.

[6] American Association for Justice. 2017. Driven to Safety: Robot Cars and the Future of Liability. , 50 pages.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* 23, 2016 (May 2016), 139–159.

[8] Jack M. Balkin. 2015. The Path of Robotics Law. *California Law Review Circuit* 6 (2015), 45–60.

[9] Jack M. Balkin. 2018. Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation. *UC Davis Law Review* 51, 3 (2018), 1149–1210.

[10] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: reorienting the study of algorithmic fairness around issues of power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 167–176.

[11] Catherine Barrett. 2019. Are the EU GDPR and the California CCPA becoming the de facto global standards for data privacy and protection? *Scitech Lawyer* 15, 3 (2019), 24–29.

[12] David Barstow. 1988. Artificial Intelligence and Software Engineering. In *Exploring Artificial Intelligence*. Elsevier, 641–670.

[13] Elena Beretta, Antonio Vetrò, Bruno Lepri, and Juan Carlos De Martin. 2021. Detecting discriminatory risk through data annotation based on Bayesian inferences. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 794–804.

[14] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 401–413.

[15] Abeba Birhane and Jelle van Dijk. 2020. Robot rights? Let's talk about human welfare instead. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 207–213.

[16] Harvard Law Review Editorial Board. 2021. Google LLC v. Oracle America, Inc. https://harvardlawreview.org/2021/11/google-llc-v-oracle-america-inc/

[17] Alexei Botchkarev. 2019. A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management* 14 (2019), 045–076.

[18] Neil E. Boudette. 2021. Tesla Says Autopilot Makes Its Cars Safer. Crash Victims Say It Kills. https://www.nytimes.com/2021/07/05/business/tesla-autopilot-lawsuits-safety.html

[19] Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible Research is Reproducible. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 725–734.

[20] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework. *European LawJournal* 13, 4 (2007), 447–468.

[21] Mark Bovens, Thomas Schillemans, and Robert E Goodin. 2014. Public Accountability. *The Oxford Handbook of Public Accountability* 1, 1 (2014), 1–22.

[22] Karen L Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[23] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.

[24] Ryan Calo. 2015. Robotics and the Lessons of Cyberlaw. *California Law Review* 103, 3 (2015), 513–563.

[25] Ryan Calo. 2021. Modeling Through. *Duke Law Journal* 72 (2021), 28 pages. https://ssrn.com/abstract=3939211 SSRN Preprint.

[26] Michael Carbin. 2019. Overparameterization: A connection between software 1.0 and software 2.0. In *3rd Summit on Advances in Programming Languages (SNAPL 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 13 pages.

[27] Danielle K. Citron and Ryan Calo. 2020. The Automated Administrative State: A Crisis of Legitimacy. , 51 pages. https://scholarship.law.bu.edu/faculty_scholarship/838 Working Paper.

[28] Danielle Keats Citron and Daniel J. Solove. 2022. Privacy Harms. *Boston University Law Review* 102 (2022), 62 pages.

[29] Ignacio Cofone and Katherine J. Strandburg. 2019. Strategic Games and Algorithmic Secrecy. *McGill Law Journal* 623 (2019), 41 pages.

[30] A. Feder Cooper and Ellen Abrams. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 46–54.

[31] A. Feder Cooper and Karen Levy. 2022. Fast or Accurate? Governing Conflicting Goals in Highly Autonomous Vehicles. *Colorado Technology Law Journal* 20 (2022).

[32] A. Feder Cooper, Karen Levy, and Christopher De Sa. 2021. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, New York, NY, USA, Article 4, 11 pages.

[33] A. Feder Cooper, Yucheng Lu, Jessica Zosa Forde, and Christopher De Sa. 2021. Hyperparameter Optimization Is Deceiving Us, and How to Stop It. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., Red Hook, NY, USA, 43 pages.

[34] Kate Crawford and Jason Schultz. 2014. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *B.C. Law Review* 55 (2014), 93–128. Issue 93.

[35] Fernando Delgado, Solon Barocas, and Karen Levy. 2022. An Uncommon Task: Participatory Design in Legal AI. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1, Article 51 (apr 2022), 23 pages. https://doi.org/10.1145/3512898

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[37] Nicholas Diakopoulos. 2020. Accountability, Transparency, and Algorithms. *The Oxford Handbook of Ethics of AI* 17, 4 (2020), 197.

[38] Finale Doshi-Velez and Been Kim. 2018. Considerations for Evaluation and Generalization in Interpretable Machine Learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven (Eds.). Springer International Publishing, 3–17.

[39] Madeleine Clare Elish. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society* (2019), 29 pages.

[40] European Parliament Committee on Legal Affairs. 2017. Report with Recommendations to the Commission on Civil Law Rules on Robotics.

[41] Federal Trade Commission. 2014. A Call For Transparency and Accountability: A Report of the Federal Trade Commission. , 110 pages.

[42] Joel Feinberg. 1968. Collective Responsibility, In Sixty-Fifth Annual Meeting of the American Philosophical Association Eastern Division. *The Journal of Philosophy* 65, 21, 674–688.

[43] Joel Feinberg. 1970. *Doing & Deserving: Essays in the Theory of Responsibility*. Princeton University Press, Princeton, NJ, USA.

[44] Joel Feinberg. 1985. Sua Culpa. In *Ethical Issues in the Use of Computers*. Princeton University Press, Princeton, NJ, USA, 102–120.

[45] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268.

[46] Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. 2016. A Confidence-Based Approach for Balancing Fairness and Accuracy. Preprint.

[47] Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael L. Littman. 2021. Model Selection's Disparate Impact in Real-World Deep Learning Applications. https://arxiv.org/abs/2104.00606

[48] Alex A. Freitas. 2014. Comprehensible Classification Models: A Position Paper. *SIGKDD Explor. Newsl.* 15, 1 (March 2014), 1–10.

[49] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347.

[50] Jeanne C Fromer. 2019. Machines as the new Oompa-Loompas: trade secrecy, the cloud, machine learning, and automation. *NYU L. Rev.* 94 (2019), 706.

[51] Harold Garfinkel. 1984. *Studies in Ethnomethodology.* Polity Press, Cambridge, UK.

[52] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[53] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster).* 11 pages.

[54] Ronan Hamon, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio Sanchez. 2021. Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY, USA, 549–559.

[55] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc., Red Hook, NY, USA.

[56] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2 ed.). Springer, USA.

[57] Deborah Hellman. 2021. Big Data and Compounding Injustice. , 18 pages. Forthcoming, SSRN preprint.

[58] Kenneth Einar Himma. 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11, 1 (2009), 19–29.

[59] Wei Hu, Zhiyuan Li, and Dingli Yu. 2020. Understanding Generalization of Deep Neural Networks Trained with Noisy Labels. In *ICLR.* 13 pages.

[60] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY, USA, 560–575.

[61] Matthias Jarke. 1998. Requirements Tracing. *Commun. ACM* 41, 12 (Dec. 1998), 32–36.

[62] Muhammad Atif Javed and Uwe Zdun. 2014. A systematic literature review of traceability approaches between software architecture and source code. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14.* ACM Press, London, England, United Kingdom, 1–10.

[63] Severin Kacianka and Alexander Pretschner. 2021. Designing Accountable Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21).* Association for Computing Machinery, New York, NY, USA, 424–437.

[64] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. arXiv:1806.02887

[65] Margot E. Kaminski. 2019. Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability. *S. Cal. L. Rev.* 92 (2019), 89 pages. Issue 1529.

[66] Sunny Seon Kang. 2020. Algorithmic accountability in public administration: The GDPR paradox. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY, USA, 32–32.

[67] Been Kim and Finale Doshi-Velez. 2021. Machine Learning Techniques for Accountability. *AI Magazine* 42, 1 (April 2021), 47–52.

[68] Alexandra Kleeman. 2016. Cooking with Chef Watson, I.B.M.'s Artificial-Intelligence App. *The New Yorker* (20 Nov. 2016).

[69] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics* 133, 1 (2018), 237–293.

[70] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5637–5664.

[71] Nitin Kohli, Renata Barreto, and Joshua A Kroll. 2018. Translation tutorial: A shared lexicon for research and practice in human-centered software systems. In *1st Conference on Fairness, Accountability, and Transparency*, Vol. 7. ACM, New York, NY, USA, 7. Tutorial.

[72] Jeff Kosseff. 2022. A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not). *Berkeley Technology Law Journal* 37 (2022), 40 pages. Issue 2.

[73] Sarah Kreps, R. Miles McCain, and Miles Brundage. 2020. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* (2020), 1–14.

[74] Joshua A. Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21).* Association for Computing Machinery, New York, NY, USA, 758–771.

[75] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165 (2017), 633–705. Issue 633.

[76] Sarah Lambdan. 2019. When Westlaw Fuels ICE Surveillance: Legal Ethics in the Era of Big Data Policing. *N.Y.U. Review of Law and Social Change* 43 (2019), 255–293. Issue 2.

[77] David Lehr and Paul Ohm. 2017. Playing with the Data: What Legal Scholars Should Learn About Machine Learning. *U.C. Davis Law Review* 51 (2017), 653–717.

[78] Mark A. Lemley and Bryan Casey. 2019. Remedies for Robots. *The University of Chicago Law Review* 86, 5 (2019), 1311–1396.

[79] Karen EC Levy. 2014. Intimate Surveillance. *Idaho L. Rev.* 51 (2014), 679.

[80] Karen EC Levy and David Merritt Johns. 2016. When open data is a Trojan Horse: The weaponization of transparency in science and governance. *Big Data & Society* 3, 1 (2016), 6 pages.

[81] Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting Inductive Biases of Pre-Trained Models. In *International Conference on Learning Representations.*

[82] Donald MacKenzie. 2001. *Mechanizing Proof: Computing, Risk, and Trust.* MIT Press, Cambridge, MA, USA.

[83] Peter Mattson, Vijay Janapa Reddi, Christine Cheng, Cody Coleman, Greg Diamos, David Kanter, Paulius Micikevicius, David Patterson, Guenther Schmuelling, Hanlin Tang, Gu-Yeon Wei, and Carole-Jean Wu. 2020. MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance. *IEEE Micro* 40, 2 (2020), 8–16. https://doi.org/10.1109/MM.2020.2974843

[84] Angelina McMillan-Major, Salomey Osei, Juan Diego Rodriguez, Pawan Sasanka Ammanamanchi, Sebastian Gehrmann, and Yacine Jernite. 2021. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. ArXiv preprint.

[85] Alexander Meinke and Matthias Hein. 2019. Towards neural networks that provably know when they don't know. ArXiv preprint.

[86] Jacob Metcalf, Emanuel Moss, Ranjit Singh, Emnet Tafese, and Elizabeth Anne Watkins. 2022. A relationship and not a thing: A relational approach to algorithmic accountability and assessment documentation. https://doi.org/10.48550/ARXIV.2203.01455

[87] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY, USA, 735–746.

[88] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, New York, NY, USA, 220–229.

[89] Pegah Moradi and Karen Levy. 2020. The Future of Work in the Age of AI: Displacement or Risk-Shifting? *Oxford Handbook of Ethics of AI* (2020), 271–287.

[90] Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, Madeleine Clare Elish, and Jacob Metcalf. 2021. Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. SSRN preprint.

[91] D.K. Mulligan and K.A. Bamberger. 2018. Saving governance-by-design. *California Law Review* 106 (June 2018), 697–784.

[92] Deirdre K. Mulligan and Kenneth A. Bamberger. 2019. Procurement As Policy: Administrative Process for Machine Learning. *Berkeley Technology Law Journal* 34 (4 Oct. 2019), 771–858.

[93] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Association for Computational Linguistics, Online, 5356–5371.

[94] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. 2017. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Red Hook, NY, USA.

[95] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2, 1 (1996), 25–42.

[96] Ngozi Okidebe. 2022. Discredited Data. Forthcoming, Cornell Law Review, Vol. 107, shared privately with the authors.

[97] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems.* Curran Associates Inc., Red Hook, NY, USA, Article 1254, 12 pages.

[98] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The Limitations of Deep Learning in Adversarial Settings. In *1st IEEE European Symposium on Security & Privacy*. IEEE, 16 pages.

[99] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48.

[100] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 8024–8035.

[101] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information" Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. ACM, New York, NY, USA, 1–21.

[102] Edward Raff. 2019. A Step toward Quantifying Independently Reproducible Machine Learning Research. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 492, 11 pages.

[103] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 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*. ACM, New York, NY, USA, 33–44.

[104] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5389–5400.

[105] Jathan Sadowski, Salomé Viljoen, and Meredith Whittaker. 2021. Everyone should decide how their digital data are used — Not just tech companies.

[106] Thomas Scanlon. 2000. *What We Owe to Each Other*. Belknap Press, Cambridge, MA, USA.

[107] Markus Schlosser. 2019. Agency. In *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, USA.

[108] Andrew D Selbst. 2021. An Institutional View Of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology* 35 (2021), 75 pages. Issue 117.

[109] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68.

[110] Hetan Shah. 2018. Algorithmic Accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018), 6 pages.

[111] Catherine M Sharkey. 2016. Can Data Breach Claims Survive the Economic Loss Rule. *DePaul Law Review* 66 (2016), 339.

[112] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value Cards: An Educational Toolkit for Teaching Social Impacts of Machine Learning through Deliberation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 850–861.

[113] David Shoemaker. 2011. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 121, 3 (2011), 602–632.

[114] Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. 2020. Optimizer Benchmarking Needs to Account for Hyperparameter Tuning. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 9036–9045.

[115] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a Design Fix for Machine Learning. ArXiv preprint.

[116] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. 2021. A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability. ArXiv preprint.

[117] Brian Cantwell Smith. 1985. The Limits of Correctness. *SIGCAS Comput. Soc.* 14,15, 1,2,3,4 (Jan. 1985), 18–26.

[118] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958.

[119] Luke Stark and Jevan Hutson. 2021. Physiognomic Artificial Intelligence. SSRN preprint.

[120] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640.

[121] Harry Surden and Mary-Anne Williams. 2016. Technological Opacity, Predictability, and Self-Driving Cars. *Cardozo Law Review* 38 (2016), 121–181.

[122] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR (Poster)*. 10 pages.

[123] Matthew Talbert. 2019. Moral Responsibility. In *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University, USA.

[124] Piotr Tereszkiewicz. 2018. Digital platforms: regulation and liability in the EU law. *European Review of Private Law* 26, 6 (2018), 18 pages.

[125] Sherry Turkle. 2005. *The second self: Computers and the human spirit*. MIT Press, Cambridge, MA, USA.

[126] U.S. Government Accountability Office. 2021. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. , 112 pages. https://www.gao.gov/assets/gao-21-519sp.pdf

[127] Kees-Jan van Dorp. 2002. Tracking and tracing: a structure for development and contemporary practices. *Logistics Information Management* 15, 1 (March 2002), 24–33.

[128] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, New York, NY, USA, 1–9.

[129] Carissa Véliz. 2021. Moral zombies: why algorithms are not moral agents. *AI & SOCIETY* 36 (2021), 11 pages.

[130] Salomé Viljoen. 2021. A Relational Theory of Data Governance. *Yale Law Journal* 131, 2 (2021), 82 pages.

[131] Sandra Wachter and Brent Mittelstadt. 2019. A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review* 2 (2019), 130 pages.

[132] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, explainable, and accountable AI for robotics. *Science (Robotics)* 2, 6 (2017).

[133] Ari Ezra Waldman. 2019. Power, Process, and Automated Decision-Making. *Fordham Law Review* 88 (2019), 21 pages.

[134] Yilun Wang and Michal Kosinski. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114 (02 2018), 246–257.

[135] Gary Watson. 1996. Two Faces of Responsibility. *Philosophical Topics* 24, 2 (1996), 227–248.

[136] Jan Whittington, Ryan Calo, Mike Simon, Jesse Woo, Meg Young, and Peter Schmiedeskamp. 2015. Push, pull, and spill: A transdisciplinary case study in municipal open government. *Berkeley Technology Law Journal* 30, 3 (2015), 1899–1966.

[137] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. ACM, New York, NY, USA, 1–18.

[138] Xiaolin Wu and Xi Zhang. 2016. Automated Inference on Criminality using Face Images. ArXiv preprint.

[139] Yichen Yang and Martin Rinard. 2019. Correctness Verification of Neural Networks. NeurIPS 2019 Workshop on Machine Learning with Guarantees.

[140] Meg Young, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2019. Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 191–200.

[141] Du Zhang and Jeffrey JP Tsai. 2003. Machine learning and software engineering. *Software Quality Journal* 11, 2 (2003), 87–119.

[142] Ruqi Zhang, A. Feder Cooper, and Christopher M De Sa. 2020. Asymptotically Optimal Exact Minibatch Metropolis-Hastings. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, 19500–19510.