Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts

JACOB METCALF*, Data & Society Research Institute
EMANUEL MOSS, Data & Society Research Institute and CUNY Graduate Center
ELIZABETH ANNE WATKINS, Princeton Center for Information Technology Policy and Data & Society Research Institute
RANJIT SINGH, Data & Society Research Institute
MADELEINE CLARE ELISH, Data & Society Research Institute

Algorithmic impact assessments (AIAs) are an emergent form of accountability for entities that build and deploy automated decision-support systems. These are modeled after impact assessments in other domains. Our study of the history of impact assessments shows that "impacts" are an evaluative construct that enable institutions to identify and ameliorate harms experienced because of a policy decision or system. Every domain has different expectations and norms about what constitutes impacts and harms, how potential harms are rendered as the impacts of a particular undertaking, who is responsible for conducting that assessment, and who has the authority to act on the impact assessment to demand changes to that undertaking. By examining proposals for AIAs in relation to other domains, we find that there is a distinct risk of constructing algorithmic impacts as organizationally understandable metrics that are nonetheless inappropriately distant from the harms experienced by people, and which fall short of building the relationships required for effective accountability. To address this challenge of algorithmic accountability, and as impact assessments become a commonplace process for evaluating harms, the FAccT community should A) understand impacts as objects constructed for evaluative purposes, B) attempt to construct impacts as close as possible to actual harms, and C) recognize that accountability governance requires the input of various types of expertise and affected communities. We conclude with lessons for assembling cross-expertise consensus for the co-construction of impacts and to build robust accountability relationships.

CCS Concepts: • **Social and professional topics** → **Computing / technology policy**; *Technology audits*; • **Human-centered computing** → *HCI design and evaluation methods.*

Additional Key Words and Phrases: algorithmic impact assessment, impact, harm, accountability, governance

1 INTRODUCTION

Algorithmic Impact Assessments (AIAs) are emerging governance practices for delineating accountability, rendering visible the harms caused by algorithmic systems, and ensuring practical steps are taken to ameliorate those harms. Multiple national governments, scholars, technology companies, and advocacy groups, have proposed mechanisms for AIAs that encompass a wide array of purposes, methods, and pathways to accountability [31, 56, 72, 91, 92, 97]. Through

*All authors contributed equally to this research.

Authors' addresses: Jacob Metcalf, jake.metcalf@datasociety.net, Data & Society Research Institute, New York, NY; Emanuel Moss, emanuel@datasociety.net, Data & Society Research Institute, New York, NY , CUNY Graduate Center, New York, NY; Elizabeth Anne Watkins, ewatkins@datasociety.net, Princeton Center for Information Technology Policy, Princeton, NJ , Data & Society Research Institute, New York, NY; Ranjit Singh, ranjit@datasociety.net, Data & Society Research Institute, New York, NY; Madeleine Clare Elish, mcelish@datasociety.net, Data & Society Research Institute, New York, NY.

our comparative study of the history of impact assessment and the algorithmic accountability literature, it becomes clear that the ***impacts* at the center of AIAs are *constructs* that act as *proxies* for the often conceptually distinct sociomaterial *harms* algorithmic systems may produce**. "Harms" only become "impacts" in an accountability relationship that obligates the designers, operators, and maintainers of algorithmic systems to identify, explain, and justify or ameliorate actual or potential harms of such systems. What counts as an adequate assessment, when that assessment happens, and how stakeholders are made accountable to each other are contested outcomes shaped by fraught power relationships [81]. The challenge we pose in this paper is how to develop AIAs as a governance mechanism while ensuring that the evaluative construct of "impact" remains close to the sociomaterial harms that need to be prevented.

A history of impact assessments from other domains, explored below, shows that **"impacts" are made knowable through highly contested and widely variable governance practices, technical requirements, regulations, and documentation**. Fairness, Accountability and Transparency (FAccT) scholarship on algorithmic systems has tended to focus on the technical metrics and intra-organizational accountability practices that will necessarily contribute to AIAs [1, 13, 22, 43, 48, 71, 93], but these measures do not yet fully constitute the accountability relationships we explore in this paper. Engaging with long standing concerns in the field of science, technology, and society studies (STS), we explore how "impacts" are co-constructed with accountability relationships [10, 11, 70]. Impact assessments of an algorithmic system do not produce accountability unless the methods used to determine impacts are submitted to a forum that has the ability to mandate changes in the implementation of sociotechnical systems (or provide remedy for harms) [15, 106]. These relationships—between algorithmic systems and a forum empowered to demand changes to such systems—are fundamental to establishing accountability in governing the impacts of algorithmic systems.

Furthermore, algorithmic impact assessments do not produce accountability unless they attend to the "constructedness" of impacts themselves. Indeed, not only must there be a relationship between actors and fora in which the forum can pass judgement on or mandate changes for an actor, there must also be a relationship in which impacts, as evaluative measures, are brought into as close an alignment as possible to potential harms. Well-established examples from FAccT scholarship show that many cases of harmful algorithmic systems are enabled by the proxy gap between formal metric and actual harm, and that such abstractions can obscure the complex relationship between social and technical systems [98]. As governments, companies and institutions begin to solidify AIA practices as an obligatory passage point for building and integrating algorithmic systems [33], mapping impacts to actual harms is not only critically important to securing ideals of fairness, transparency, and justice, but also necessary for accountability. As we will demonstrate, the constructedness of impacts can lead to a thin, attenuated form of accountability that satisfies formal guidelines for the operation of algorithmic systems but fails to ensure that an adequate array of potential "real world" harms is part of that evaluative framework.

This paper will examine recent proposals for algorithmic impact assessment and existing impact assessment processes from other domains to show how existing proposals do not yet structure relationships in ways that produce accountability, nor do they adequately interrogate the relationship between expected outcomes measured as algorithmic impacts and the potential harms algorithmic systems produce. The paper will conclude with pragmatic recommendations for algorithmic impact assessment to address these challenges. Throughout, we argue that *algorithmic impact assessments must map impacts rigorously to potential harms, determine that an adequate scope of potential harms has been rendered measurable as impacts, and produce accountability by distributing power between differently-positioned stakeholders.*

2 ACCOUNTABILITY AND THE ASSESSMENT OF "IMPACT"

Impact assessments are a particular governance practice that structures and sustains an accountability relationship, and have been widely established in domains analogous to algorithmic systems [51]. Following a definition offered by Mark Bovens, accountability (in general) requires: 1) an actor, who submits an often technical account of the impact of a proposed/implemented system 2) a forum that evaluates this account and can propose changes in the system's implementation based on the evaluation of its impact 3) the structured relationship between the two, 4) the content and criteria for accounting impact, and 5) the consequences arising from the account [15]. Drawing on Bovens, Maranke Wieringa developed a definition of algorithmic accountability [106]:

> Algorithmic accountability concerns a networked account for a socio-technical algorithmic system, following the various stages of the system's lifecycle. In this accountability relationship, multiple actors (e.g. decision makers, developers, users) have the obligation to explain and justify their use, design, and/or decisions of/concerning the system and the subsequent effects of that conduct. As different kinds of actors are in play during the life of the system, they may be held to account by various types of fora (e.g. internal/external to the organization, formal/informal), either for particular aspects of the system (i.e. a modular account) or for the entirety of the system (i.e. an integral account). Such fora must be able to pose questions and pass judgement, after which one or several actors may face consequences. The relationship(s) between forum/fora and actor(s) departs from a particular perspective on accountability. (p.10)

Although there are potentially many forms of accountability in any domain, impact assessments are an established and fairly direct way to construct an accountability regime. As proponents of AIAs develop and test viable proposals for, and subcomponents of, future AIAs, there is a demonstrable variety of perspectives on who is the actor, who is the forum, what is their relationship, what needs to be reported, and what are the consequences. As we will show below, every domain has a different (and historically contingent) set of expectations about **what constitutes impacts and harms** within that domain, **how those potential harms are best assessed as the impacts** of a particular undertaking, **who is responsible for conducting that assessment**, and **who has the authority to demand changes to that undertaking**. When comparing impact assessments across domains, it becomes clear that "impacts" are not directly observed or measured. Indeed, accounting for an "impact" often requires positing a counterfactual situation wherein the proposed action (e.g., deploying a technology, building a road, spending a budget, etc.) does not happen, which makes the assessment of impacts as much a thought experiment as an empirical endeavor.

Understanding impacts therefore requires attention to how such proposed actions can be assessed. Briefly, we are informed by science and technology studies perspectives that show the deep concomitant relationship of scientific practices with their social and political environments. Described variously by terms such as co-production [53], co-construction [68], agential realism [5], actor-network theory [63], and mutual shaping [9], this shared analytic framework articulates how scientific practices that produce knowledge emerge through, and are inextricably intertwined with, the social conditions that render them necessary and possible. We will use the term co-construction generically, although there are subtle differences in this scholarship. The central theme here is that *scientific objects* are not plainly present to us in a pre-interpreted state, and that defining and measuring scientific practices depends upon contingent social conditions, institutions, consensus building, and norms which are not themselves "objective." Conversely, social and political conditions are also produced through scientific practices that set (and constantly change) the boundaries of what can be known and acted upon. Therefore co-construction is the process by which social and scientific conditions

mutually shape each other. Crucially, *constructed* in this sense is not in opposition to *real*, rather it is a description of how a phenomenon becomes real and scientifically legible.

Impacts are co-constructed objects that emerge through the negotiation of accountability relationships. As actors and fora enter into contestations over what an undertaking is and what it does in the world, ways of describing and evaluating its effects must be rendered mutually legible between these differently-positioned stakeholders. By agreeing upon categories of "impacts", these stakeholders stabilize impacts as evaluative objects upon which they can act. The *what* of an algorithmic impact is co-produced with the *who, when, where* and *why* of algorithmic accountability. However, these impacts exist at a different level of abstraction [98] from the harms that an undertaking may produce in the world, they are proxies for *harms* that are convenient to use within relationships of accountability but that must constantly be scrutinized to ensure that they are adequate and appropriate *proxies* for "real-world" harms.

Therefore, we propose re-purpose Wieringa's definition of algorithmic accountability by asserting that in the context of impact assessment actors and fora also have an obligation to explain and justify to each other the ways by which they construct their accounts of the socio-technical algorithmic system itself—in part (modular) and in total (integral). Evaluative constructs cannot simply be adopted from existing audit or measurement techniques without critical scrutiny. Instead they must be examined as to their fitness for purpose. Similarly, evaluative constructs cannot be produced through disciplinary silos, and any account produced from one disciplinary perspective should be cross-examined by complementary methods to triangulate on actual or potential harms. Finally, evaluative constructs cannot be declared by fiat. Any prescriptive declaration of what constitutes an adequate account of a socio-technical algorithmic system is in peril of eliding actual or potential harms that might be missed by a prescribed process. Rather, such accounts must be subject to a deliberative, consensus-based process that stays as close to potential harms by privileging justice and dignity for those most likely to be affected by such harms.

2.1 Impact Assessment in non-algorithmic contexts

The history of impact assessments in other domains [4, 8, 24, 26, 31, 37, 41, 42, 54, 60, 64, 65, 78, 86, 87, 103] shows how they have evolved as an accountability mechanism by leveraging impacts as evaluative constructs made tractable through **organizational, legal, political, and epistemic contestations**. This history illustrates a complicated, contested, and fraught set of mechanisms to set up accountability relationships both in terms of how such relationships are structured and how they map impacts to actual harms. The terms of reference used for "harms" in each of these domains are different, for example, tort law in various jurisdictions have particular definitions of harm. We take a broader conception of harm in this paper as experienced in the real world. As Bovens and Wieringa both indicate in their descriptions of accountability, without a forum that can mandate changes to an undertaking, whether it be a housing development, a gas pipeline, a logistics supply chain, or a database, the accountability process cannot mitigate or ameliorate the potential harms produced by the undertaking. Furthermore, without an adequate description of an undertaking's technical details, it is impossible to determine whether the anticipated harms and benefits can reasonably be expected.

2.1.1 Assessing Impacts. Many established forms of impact assessment have a roster of well-developed techniques and methods that can be applied to particular types of projects, as circumstances dictate.[1] Impact assessment methodologies are primarily concerned with *measurement*, particularly of how a project produces impacts that diverge from a baseline. Impact assessments strive to determine what the impacts of a project are *relative to a counterfactual world* in which that project does not take place or in which an alternative project takes place. Therefore, an EIA assesses impacts to a water

[1] See https://iaia.org/best-practice.php for an in-depth selection of IA methods.

resource by estimating what the level of pollutants is likely to be, as compared to what the level of pollutants otherwise would be [78]. A human rights impact assessment (HRIA) will document the impact to specific human rights relative to what the human rights landscape already looked like in a particular jurisdiction. Although humans are often *harmed* by human right violations, HRIAs evaluate *impact* on human rights as abstract conditions of securing life chances within a jurisdiction [60]. And a fiscal impact assessment (FIA) will assess what a municipality's fiscal situation will look after a development is completed, compared to what it would have looked like had that development not taken place [23, 61].

Environmental Impact Assessments (EIA), for example, construct "impacts" as effects on water, air, soil resources, etc. caused by an undertaking, even though the harms they intend to mitigate are to human health and the future ability to utilize those resources. In the U.S., the National Environmental Protection Act empowers state and federal agencies to demand private developers make changes to undertakings that are determined to have outsized negative impacts on the environment [30]. EIAs assess impacts to environmental resources, conceptualizing these resources as a form of service or support to a community. Impacts measured, in this domain, are defined as changes to the ready availability and viability of environmental resources. However, the harms that such impacts might cause are not explicitly assessed through measurements of impacts to environmental resources considered. Environmental destruction, decreased enjoyment from a loss of access to clean waterways, diseases either catalyzed or worsened through exposure to polluted water, food crops rendered inedible—these harms are felt in the rhythms of everyday life, yet are not measured via the way that impacts are operationalized in EIA processes.

2.1.2 Structuring Accountability and Expertise. An important feature of how impacts are determined within IA processes in other domains is how diverse forms of expertise are assembled. EIAs, for example, employ wildlife biologists, fluvial geomorphologists, archaeologists, architectural historians, ethnographers, chemists, and many others to assess the panoply of impacts a single project may have on environmental resources [79]. FIAs require fewer personnel to complete, but draw on a wide repertoire of assessment techniques that follow the same pattern as EIAs. The more varied the types of methods employed in an assessment process, the wider the range of impacts that can be assessed. Furthermore, disciplinary expertise must often be combined in order to construct the relationship between "impact" and "harm"—chemists' expertise on how toxins move through groundwater, clinicians' knowledge of how that toxin affects bodies, developmental psychologists' knowledge of how children who have been exposed to the toxin learn—all trace a path from that which is measurable as an impact from an environmental project to a harm (or set of harms) experienced by a person or community.

Impacts—and impact assessments—function as boundary objects; they have specific meanings for experts within disciplines, but are malleable enough to hold their meaning across disciplines and become productive sites of collaboration [101]. The environmental impacts mean something different to a wildlife biologist as they do to a soil scientist, but both experts can discuss "impacts" in mutually recognizable ways. Precisely *how* these various forms of expertise are brought into relation with each other, relative to the institutions playing the role of actor and forum in an accountability framework, is crucial. Some of the expertise needed to assess the impacts of a project must necessarily be provided by the actor responsible for furnishing a technical description of the project being assessed. The forum also needs specific types of expertise to evaluate whether the anticipated impacts are reasonable given the project being reviewed. This leaves a gap in the range of expert knowledge practices, necessary to evaluate impacts but are often not found within the types of organizations (private companies) that initiate projects or the types of institutions (often, but not always, government agencies) that act as fora for impact assessment processes. In many domains, consulting companies fill

this gap by providing human rights experts to conduct field interviews for HRIAs, archaeologists to conduct cultural resource surveys for EIAs, or urban planners to conduct FIAs.

The structural relationship between the actor and the forum is crucial for how an IA process produces accountability. EIAs separate development agencies or private companies who undertake projects from regulatory agencies responsible for passing judgement on environmental impact. However, not all domains' impact assessment processes delineate the "actor" and the "forum" as separate entities, especially when an IA process is not mandated by administrative law. For an HRIA, a company is *itself* accountable for its impacts to human rights by commissioning an impact assessment, but also acts as its own accountability forum when it decides which impacts it chooses to address, and how. Formally, the company acts as both actor and forum within the HRIA framework (in some cases, different organizational units with the company may be actor and forum). However, proponents of HRIAs would argue that the public sphere acts as an additional forum that can pass judgement on the corporation through public censure, boycott, or other reputational harms, which the corporation is thought to have an interest in avoiding. Similarly, Privacy Impact Assessments (PIAs) task agencies with assessing their own privacy impacts, and asserting that those assessments are adequate [82]. The agency acts as its own forum, even if the agency may face applicable fines under other laws and regulations, or may face reputational harm or civil penalties for the material privacy harms (breaches, etc.) it fails to protect against.

When an actor acts as its own forum, it creates conditions for what Lauren B. Edelman and Shauhin A. Talesh call "legal endogeneity," in which organizations construct the meaning of a law or regulation for themselves [35]. Edelman and Talesh point to the distance between the law's need for ambiguity and business's need for managerial discretion as the gap through which companies begin to set their own terms for compliance with regulation. In the context of impact assessments, private companies that act as their own forum of accountability may, through a process of legal endogeneity, eventually set their own definitions and standards for algorithmic impacts. Similarly, they may define impact assessment for themselves as a checkbox item that only asserts whether they have or have not assessed algorithmic impacts, without describing the design of a system or its potential impacts for a forum that can mandate changes, let alone assign responsibility for any harms that their system produces. While self-study of these questions is certainly preferable to no study, intra-organizational impact assessments lack the external pressure that is often necessary for a fully developed accountability relationship.

2.2 Impact Assessments emerging in algorithmic contexts

Existing proposals for AIAs differ in how they construct accountability relationships. As evident in other domains, AIAs will need to address the actor (who), the fora (when and where), and the content (what) to create effective algorithmic accountability regimes. Extant and proposed AIAs take a wide variety of perspectives on these questions, and there is not yet a coherent picture of how to co-construct both the *what* of impacts or the *who, when,* and *where*.

Andrew Selbst, in his examination of governance for big data policing, argues that a critical aspect of AIAs is simply building the capacity to describe what a proposed system actually does [97]. He notes that the vagaries of software procurement means that government agencies may deploy algorithmic systems without actually understanding what they do, which can lead to harms such as disparate impact in policing. He observes that AIAs now have primary roles similar to the early days of other IA efforts: requiring an actor (whether designer, procurer or deployer) to demonstrate early consideration of different options and the resulting externalities. Indeed, the lack of public transparency into how consequential automated decision systems operate is thus far a key driving force behind advocacy for AIAs, which have focused more directly on inquiring how a system works and left questions about the accountability relationship largely implicit.

AI Now's 2018 proposal for AIA's—an early entrant into the ongoing conversation—describes how public agencies could utilize extant procurement oversight to govern algorithmic decision systems (ADS) [92].[2] Algorithmic systems pose peculiar challenges for procurement oversight, because on the one hand they appear to be functionally similar to other enterprise software systems that are handled routinely, but on the other hand behave quite differently. ADSs may process sensitive data in an unexpected and "black-boxed" manner that is not appropriate for public agencies accountable to the public interest. AI Now's proposal constructs AIAs as an accountability relationship between public agencies and the public that by extension, forces vendors to be transparent about the workings of their ADS and make them available for public scrutiny. Some similar initiatives have been taken up at the municipal level, such as the register for algorithms used in public services in Helsinki and Amsterdam [55].

Similar to AI Now's proposal are the AIA requirements recently instituted by Canada's Treasury Board, an oversight agency mandated to provide guidance to other agencies on responsible procurement of ADSs. Their directive requires any government agency or vendor serving a government agency utilizing ADS provide an AIA, defined as "A framework to help institutions better understand and reduce the risks associated with Automated Decision Systems and to provide the appropriate governance, oversight and reporting/audit requirements that best match the type of application being designed" [85]. The Canadian AIA is described by one of its architects, policy advisor Michael Karlin [57, 58], as an electronic survey which helps institutions "evaluate the impact of automated decision-support systems including ethical and legal issues," which assigns numerical scores in a rubric format to identify risk tiers. It is contained in a Github repository to allow developers to place it close to their technical workflows [84]. Among the questions posed are: "Are stakes of the decisions very high?"; "Are the impacts resulting from the decision reversible?"; "Is the project subject to extensive public scrutiny (eg: due to privacy concerns) and/or frequent litigation?"; and ""Have you assigned accountability in your institution for the design, development, maintenance, and improvement of the system?" [83]. As Canadian critics of this process have pointed out [66], a scored Yes/No rubric is a particularly shallow form of accountability because it does not yet require an accounting of the workings of the ADS, what epistemic practices and metrics were used, or even how such systems are assembled. Without requiring demonstration of subject-matter expertise, public scrutiny of these systems cannot effectively serve as a fora and therefore reduces the ability to protect the most vulnerable people—at best they can enable the agency to create risk tiers when choosing between vendors.

The 2019 Algorithmic Accountability Act (AAA) proposed in the US Congress establishes a different accountability relationship by requiring all companies of a certain size that make use of data from regulated domains to conduct an AIA prior to deploying or selling their systems (and to retroactively conduct an AIA for all existing systems) [27]. The law was framed to ensure that algorithmic systems are held to the same non-discrimination standards that apply to other economic activities in regulated domains (e.g., financial loans, real estate, medicine, etc.) [14]. In this model, the public regulator requires an assessment, but does not facilitate a forum in which the assessment can be scrutinized, explicitly leaving it to the company's discretion to make the AIA public. While there may be market-based reasons to make such assessments public (e.g., earning public trust, smoothing business-to-business relationships), and such documentation would be discoverable during civil or criminal legal proceedings, the accountability relationship between the actor and fora is attenuated without public transparency.

The European Parliament has also investigated AIAs as a potential regulatory framework. While their preliminary study does not come down on a specific recommendation, they explore the effectiveness of both models discussed above [40]. As Margot Kaminski and Gianclaudio Malgieri argue, the European approach has favored a model of "collaborative

[2]We use "ADS" and "algorithmic system" interchangeably, throughout.

governance" between public and private entities, in which a public regulator requires all data processors that hold data about data subjects to provide transparency about their systems upon request [56]. They suggest that the scaffolding for AIAs in Europe is already present in the GDPR law, likely requiring only changes in administrative law. This approach relies on systemic governance—ensuring that a general transparency and enforcement regime is always already operative—rather than using individual AIAs as obligatory passage points through which other affected parties can construct an opposition to the system's deployment (like in the case of an EIA). The diversity of accountability relationships found in these different models of AIAs suggests the need for developing consensus on what the purpose of AIAs should be.

2.3 Impact Assessment Precursors & Components

In the context of these existing proposals, crucial insights and interventions into the "impacts" of algorithmic systems thus far have largely come from outside the accountability relationships that define IAs.

First, critical external audits have played a major role in both drawing public attention to harmful applications of machine learning, and thereby driven technology companies to create in-house governance mechanisms. By "critical external audit", we group together disparate methods of journalists, technicians, and social scientists who have examined the consequences of already-deployed algorithmic systems and who have no formal relationship with the institutions designing or integrating the audited systems. Perhaps most consequential has been the Gender Shades Project and resulting papers, led by Joy Buolamwini. Through a series of research papers, media projects and public presentations, Genders Shades has demonstrated how publicly available facial recognition systems sold by major tech companies were significantly more imprecise with dark complected faces in general, and most imprecise for dark complected women's faces in particular [22]. The result of this algorithmic bias was potentially unfair treatment by institutions using these APIs, including governmental agencies with law enforcement powers. In a follow-up paper a year later, Inioluwa Deborah Raji and Buolamwini examine the effects of their first critical audit and compared the targets from the first Gender Shades study to a control group of non-target APIs. They showed that APIs which had been previously audited were significantly more less biased than those which were not previously audited, indicating that external critical audits can result in meaningful pressure for change in organizational and technical practices of targeted companies [21, 90]. In public discussions of their research, Raji has argued that while such audits can be successful at forcing technology companies to change their practices one at a time, their most significant impact could be in forcing the creation of general obligations to conduct full fledged impact assessments [34]. Other notable external critical audits include ProPublica's examination of Northpointe's recidivism prediction API [3], and the Allegheny County, Pennsylvania, Office of Children, Youth and Families' machinic prediction of child abuse risk [104], which received technical [25], ethical [32], and social scientific [38] audits.

While external audits have illustrated the downstream consequences of algorithmic systems, internal governance mechanisms have played an important role in showing how technical and organizational practices are essential components of assessing impacts. Notably, Google research teams have developed a series of governance mechanisms integrated with the core engineering processes of machine learning product development: (1) Model cards: Records of how models have been developed which are portable with the model, resolving the challenge of inadvertently repurposing a model in a risky manner when it is put to use by another development team or made open source [76]. (2) Datasheets: Documents that capture the conditions of learning sets, enabling important engineering and ethical context to be portable with the datasets that form the foundation of machine learning applications [45]. (3) End-to-end internal

accountability: Integrating these mechanisms in maintaining lifecycle internal accountability can provide organizations with a series of documents that facilitate governance of the systems they develop [91].

These internal mechanisms and documents are critical to integrating ethics governance with established software engineering practices and technical tools. Whatever form AIAs take, internal governance such as model cards and datasheets will undoubtedly make fulfilling the requirements of AIAs far more efficient and thorough. Nonetheless, internal governance will always run the risk of legal endogeneity and lack external fora that can demand accountability for harms. Similarly, external critical audits, wherein the auditor has no formal access to the internal workings of the system, lack the ability to render impacts as changes to the system.

2.4 Tracing the Use of "Impact" in FAccT

The challenge at hand for defining algorithmic "impact" that accountability mechanisms can act upon is how to render harms visible in the routine focus on measurable outcomes.

The notion of "impact" has been central to discussions of algorithmic fairness from the earliest workshops and conferences on the topic [94]. The most voluble discussions of impact have focused on the "disparate impact" of algorithmic systems, in which such systems replicate, extend, or entrench systematic discriminatory social biases in the classifications they make [6, 39]. Disparate impact describes uneven outcomes of an algorithmic system, which may not be intentional on the part of anyone involved in the development or operation of that system. It stands in contrast to the "disparate intent" of discrimination that arises from purposeful decisions made by those involved in a decision-making process [109]. Disparate impacts are part of a larger set of automated systems' unintended consequences, the full complement of which Oscar Gandy argues "should be subject to routine, if not continuous assessment and regulatory control" [44]. In subsequent years, the FAccT community has undertaken an intensive research program into algorithmic fairness and many of the discriminatory impacts of algorithmic systems that would be necessary to assess and regulate such systems, as Gandy suggested in 2010. However, the relationship between the "impacts" that have been studied by the FAccT community and the full complement of algorithmic systems' unintended consequences has not been as thoroughly interrogated.

As momentum builds for algorithmic impact assessments (AIAs) such as those laid out in several promising frameworks [56, 91, 92, 97], to serve as a mechanism for addressing the unintended consequences of algorithmic systems, the question of what constitutes an "impact" in the context of an AIA has never been more important. Distinguishing between, on one hand, impacts as evaluative constructs that describe the unintended consequences of algorithmic systems in ways that make them amenable to assessment and regulatory control, and on the other hand, the unintended consequences themselves as the concrete potential harms that individuals and groups may experience through the operation of these systems, is key to this task.

The distance between "impacts" as measures of the difference in probabilities of a classificatory outcome between demographic categories, and the tangible risks and harms that arise from that difference is at stake here. While computational methods have demonstrated the ability to describe the former (disparate classificatory probabilities) in any number of important ways [1, 12, 43, 105], computational methods are less well-suited to measure the ways algorithmic systems produce and distribute the risk that people and groups might experience as a result of these classificatory processes. Risk has a productive life, in Caitlin Zaloom's framing [108], that generates value for those who can distribute it across society, and this is increasingly accomplished through machine learning [80]. However, the practical effects of the distribution of risk—how it leads to negative consequences for people and groups—often require other forms of expertise and knowledge (including local and indigeneous knowledge).

In the absence of these forms of expertise and knowledge, efforts to address only the measures of potential impact will, per Goodhart's Law [46], fail to minimize the tangible harms to people. Bureaucratic legibility demands that the qualitative, lived experience of harms be transformed into quantified, comparable impacts [36, 88]. Rendering harms as impacts is an exercise that is both the product and reproduction of power. It is no accident that those who suffer the most are the least likely to be captured by impacts and wield the least power within, over, or through IA processes, a pattern that has been well-studied in the context environmental impacts and environmental racism [20, 28, 107]. Furthermore, IAs also threaten to erase, through presumptions of comprehensiveness, other kinds of harm being done. The presumption of comprehensive representation of risks and harms through impact assessments can be dangerous. Risk assessments in the domain of environmental protections, for example, have played a role in setting allowable exposure limits that have ultimately been found to be harmful to children, for example in recent findings that once-permissible levels of lead exposure actually led to cognitive and behavioral effects [77].

In the context of algorithmic fairness, transparency, and accountability, computational methods for evaluating impact do not themselves constitute accountability. Rather, computational methods are resources through which accounts of a socio-technical algorithmic system (to echo Wieringa's definition of accountability) can be constructed and submitted to a forum. However, these methods will remain partial, unless they are able to connect the operation of ADSs to the potential harms of such systems. "Impacts" can gather the necessary accountability relationships to resolve actual harms to people only by taking into account (1) the conceptual groundwork for defining them in the prior history of IAs from other domains, existing proposals for AIAs, and extant governance mechanisms; and (2) the challenges for assembling a governance regime around AIAs in the future. We contend that accomplishing this will require approaching impacts as co-constructed with accountability relationships.

3 ASSESSING "IMPACTS", ELIDING "HARMS"

Whether in non-algorithmic or algorithmic domains, establishing the *"what?"* of impacts is always already an exercise in establishing an accountability relationship. Assessing impacts is not a *revealing* of harm, it is a *constructing* of proxies for harm in a context where accountable parties have agreed to the *who*, *when*, and *how* of preventing or addressing harms. We have argued that current AIA proposals, and supporting internal governance practices, do not yet accomplish this task. We have also argued that IA practices generally suffer from a widely recognized problem in FAccT scholarship: the need to quantify the effects of algorithmic systems in order to act upon them leads to an ontological flattening that misses the actual lived harms in favor of actionable proxies. Even in cases where harms are identified, they may not compute within the parameters of impact assessment, and fall short of providing justice to affected populations. In this section we show how this gap has played out in two prominent examples of attempts to assess the impacts of algorithmic systems.

In 2018 Facebook partnered with the consulting firm Business for Social Responsibility, to conduct a HRIA on its role in the conflict in Myanmar, a country "plagued by political and social divisions" [102]. HRIAs focus on impacts to human rights as enumerated in the Universal Declaration of Human Rights, later codified into the International Bill of Human Rights (IBHR) [60], including freedom from torture, freedom of expression, and right to livelihood, among others. These impacts are not constructed as harms to individual rights holders. The HRIA report, faithfully following the guidelines as laid out in the UDHR and the IBHR as well as the UN Guiding Principles on Business and Human Rights, identified that Facebook's activities had "actual impact" on human rights such as, among others, security, privacy, and freedom of expression, with, for example, one impact on the human right to security: "Accounts being used to spread hate speech, incite violence, or coordinate harm may not be identified and removed" [19]. This report did not describe the harm to

actual humans that were a direct result of Facebook's platform governance choices, wherein it has been accused of playing a "key role" in a conflict where 650,000 Rohingya refugees were forced to flee to Bangladesh [50], and the UN Human Rights Chief strongly suspected acts of genocide. Genocide was mentioned exactly two times in the report, once to highlight actions that Facebook had taken to remove military officials from the platform—described as a "strong statement" to ward off potential genocidal activities—and again to highlight the positive role that Facebook could play providing remedies for victims of genocide [75]. The refugee crisis that Facebook was accused of contributing to via its affordances that enabled hate speech and propaganda went unmentioned in the impact assessment of the very incident that prompted the need for this internal study in the first place.

Compounding these concerns are the structures—or lack thereof —which HRIAs establish to ensure that they function as documents of accountability. Questions around who is responsible for conducting these assessments, and who has the authority to demand changes to the relevant undertaking, go unanswered (starkly so in the case of Myanmar). HRIAs are under no requirement to be conducted by, or include oversight from, a governing body, or even include the input of any impacted population or the body politic. They contain no mechanisms which can be enforced through accountability forums such as legislative, judicial, or electoral backstops. In fact, HRIAs are typically conducted after an event has taken place, entirely eliminating any potential of preventing the very harms they describe. Further, once produced by a third-party or independent firm, these documents usually are distributed only internally, as there are no requirements for public visibility or transparency. Unusually, the study conducted by BSR on Facebook's role in Myanmar was published, and the company's activities in the country did lead to what's been described as a "collapse of public trust [and] … heightened scrutiny from lawmakers" [62], but without an external accountability relationship there have been no firm consequences for Facebook.

For our second example, we move onto the Allegheny CYF risk scoring algorithm, which is one of the most thoroughly studied algorithmic systems in use by a public agency. Although it has not been subject specifically to an AIA (since no mandate to do so applies in its jurisdiction), it has undergone multiple audits by a variety of experts. In 2015 the Allegheny County, Pennsylvania, Office of Children, Youth and Families (CYF) solicited proposals for a tool which could help administer public resources to children at risk of severe harm, including sexual abuse, physical abuse, and death. An automated decision-support system was built to help government screeners who processed reports of potential abuse to make decisions about which reports required in-person investigations, and to ensure that resources were efficiently directed towards children at substantial risk of suffering harm [25]. The abuse investigators themselves were firewalled from the algorithmic system so as not to influence their own determinations of harm. The system used data from public agencies to produce two predictions: first, that the child, if screened out of the process, would be referred back into the system again; and second, that if they were screened in, they were likely to be removed from their family and placed in foster care. Agencies from which data was drawn included those for mental health, criminal justice, education, and prior contact with the CYF department, and sometimes multiple generations of family members' records from these agencies were automatically tabulated in the scoring algorithm and reports.

Eubanks points out two types of harm produced by this system. First, the potential harms the tool predicted—chances that the child would be screened in again, and chances the child would be removed from their families—weren't actually harms at all, but procedural and community-determined proxies for harm [38]. In other words, the ADS doesn't look for harms, but rather, proxies which are more visible and salient to machinic processes: it uses data that is collected and produced when an agency responds to a child being harmed. Actual harms may be missed and thus overlooked by the creation of and adherence to the predictions of this system. The second type of harm that Eubanks details is not one of blindspots created by classification, but the hazards of classification itself, when such classification measures are

intrinsically and systematically biased along lines of social and economic vulnerability: "the activity that introduces the most racial bias into the system is the very way the model defines maltreatment" (p. 155) [38], because it uses data measured by an unfair system to predict what that same unfair system would do in similar situations. In other words, harms are not just missed but actually created through the epistemic circularity of machine learning applications applied to historical records of human behavior [16]. The design of AIAs must anticipate and mitigate the issues of "constructedness" in both systems. If AIAs were to uncritically render harmful classificatory schema as "neutral" metrics by which to measure impacts, it would threaten to produce layered proxy-reliant processes obscuring the experiences of the public they purport to serve.

4 CO-CONSTRUCTING ALGORITHMIC IMPACT

Assessing impacts requires rendering sociomaterial harms as evaluative objects for institutional action [95]. Impact assessments will not be a sufficient way of both measuring and ameliorating the material harms that ADSs present for publics unless impact assessments, as a form of administrative accounting, adequately *represent the harms that are actually potential outcomes of a system.* But, as we have argued, the deck is stacked against doing this easily with algorithmic impacts: the harms of algorithmic systems are dispersed and aggregate; the technical and organizational practices of the tech industry militate in favor of strictly numerical metrics; and there are already many examples of algorithmic harm that derive from the gap between actual lived experience and machinic proxies. The question then is: how can an effective governance regime *that actually addresses harms* be forged around such a constructed object?

Perhaps counter-intuitively, we argue that it is the co-construction of "impacts" that provides the guideposts to an effective form of algorithmic governance. Because there is no interpretation-free access point to "impacts", it then becomes necessary to carefully attend to how those impacts get constructed.

4.1 Understanding the Assessor's Regress

Since impacts are an evaluative construct designed to minimize harms, establishing a concrete relationship between impacts and harms can become a recursive problem. There is no way to define impacts without harms, but there is also no way to delimit harms without the affordances of the impact assessment process.

This recursion can be illustrated as follows:

Q : How do we detect algorithmic harms?
A : We conduct an AIA to assess the likely impact of an algorithmic system to people who may experience these harms.
Q : How do we know if the AIA is assessing algorithmic harms to people who may experience it adequately?
A : We know that the assessment is adequate if the AIA detects possible harms caused by the algorithmic system.
Q : How do we detect these possible harms?
A : We conduct an AIA.
And so on...

As the relationship between impacts and harms is interrogated, it becomes more tightly coupled and a more complete set of harms is rendered legible as measurable impacts. And yet, absolute certainty that the full complement of harms has been rendered legible remains forever elusive. This is what we name "the assessor's regress": the completeness of the assessment relies on a never-ending chain of justification. Since organizations responsible for algorithmic systems can only be held accountable for the impacts that can be foreseen by IA processes and not the full set of possible

outcomes that an algorithmic system might produce, this regress becomes both the means of ensuring accountability but also an organizational challenge as it becomes difficult to know when a satisfactory array of harms have been rendered legible.

This type of epistemic regress is an established phenomenon in STS literatures. Harry Collins has identified a similar regress in studies of scientific controversies to theorize the experimenter's regress: scientific facts can only gain legitimacy if they are produced by legitimate instruments and at the same time, instruments gain legitimacy if they produce legitimate scientific facts [29]. Along similar lines, Donald MacKenzie has conceptualized the tester's regress: "If there is no agreement as to what constitutes successful testing, and what the correct results [for evaluating performance during a test] are, then it becomes impossible independently to evaluate competence" (p. 414) [69] in testing technologies before deploying them. This regress has profound consequences for the processes of establishing accountability.

Grappling with these consequences, Michael Power has argued that such regress can engender a broader mistrust in allocating accountability: "If those engaged in everyday work are not trusted, then the locus of trust shifts to the experts involved in policing them, and to forms of documentary evidence or in management assurances about system integrity [and performance]. Ultimately there is a 'regress of mistrust' in which the performances of auditors and inspectors are themselves subjected to audit" (p. 11) [89]. When institutional practices of consensus and inclusion are working well, the regress becomes less visible and the constructedness of the object of study (e.g., impacts, experimental apparatuses, testing regimes) is not an ongoing concern. When the institutional practices lack legitimacy and consensus is not actively sought and maintained, the object of study becomes far more contested.

There is no "impact" without the accountability practices that define, detect and act upon it. Likewise, there is no accountability without defining by ongoing institutional consensus on what an impact is. However, these uncertainties do not imply that conducting rigorous AIAs and establishing accountability is impossible. Resolving assessor's regress requires intentional institutional practices and assembling expertise that can build consensus over whether the adequate scope and weight of possible algorithmic harms has been measured and accounted for during AIAs. *The assessor's regress is only closed by a forum in a legitimate accountability relationship.*

4.2 Assembling Expertise

The assessor's regress makes "impacts" distinctly different types of objects of study than "harms." Yet if the purpose of engaging in the development of algorithmic accountability at all is to prevent and ameliorate harms to people and society, then impacts must map as closely as possible onto actual and potential harms. We suggest this will be a challenge for constructing "algorithmic impacts" in a way that it may not have for related forms of assessment in machine learning, as in practices of algorithmic fairness, AI safety, computational security, and digital privacy, which are deeply invested in instrumentation, measurements, and metrics that pertain to the internal performance of algorithmic systems. By internal performance, we are referring to metrics like false positive rates, precision and recall, AUC, ROC when tested against validation and holdout data, instrumentation that enables penetration testing, or measurements that enable the kinds of interpretability with which AI safety researchers are concerned [2]. While they are essential components of an assessment process, **if we start and end at impacts that can be measured using only these tools, then we will not be able to create effective algorithmic impact assessments**.

Because there are no impacts without the accountability practices that define and "detect" them, the tools that are developed to identify and evaluate impacts will shape what harms are detected. Like all research questions, what is uncovered is a function of what is asked, and what is asked is a function of who is doing the asking [49]. The algorithmic

impacts most likely to be detected today are those for which the most robust set of metrics and measures have been developed. Researchers within the domain of algorithmic fairness have generated a large corpus of work measuring disparate algorithmic classifications, non-representative datasets, and other forms of statistical bias in algorithmic systems, as well as methods for minimizing these disparities. However, this work has been narrowly focused, at least in the U.S., on disparities between demographic categories protected by Title VII [59]. In part, this represents the interests of private companies seeking to limit liability and government agencies seeking to demonstrate compliance as they employ algorithms as part of their organizational practices. But not only are these Title VII categories incomplete with respect to forms of discrimination [17], discimination is incomplete with respect to algorithmic harms. In these conditions, the legal frameworks and the technical tools for identifying and remediating algorithmic impacts suffer from a form of legal *and* technical debt that makes it difficult to understand and measure, let alone remediate, harms beyond those caused by disparate impact.

The observation that impacts are constructs shaped by existing technical and legal debt need not indict the pursuit of fair, responsible, accountable approaches to algorithmic system design. Rather, it instead creates new possibilities for thinking about how impacts, as evaluative constructs, can be shaped to more closely align with the potential harms of algorithmic systems. The distance between impacts and harms can be continuously monitored and additional potential harms can be detected and made assessable as impacts over time. To do so, however, will require that technical expertise from computer science, machine learning, and database engineering is complemented by other forms of expertise that can address a more complete set of harms.

While technical expertise in computer science is certainly necessary to audit the performance of algorithmic systems and adequately describe their operations, the formal limits of algorithmic expertise are in tension with the social worlds in which algorithmic systems operate. Borrowing from Ben Green and Salomé Viljoen's "algorithmic realism" [47], we expand the scope of algorithmic expertise to include design researchers who may be able to investigate impacts caused by harmful "dark patterns" in user interfaces [18], behavioral scientists who may be able to investigate "opportunistic… choice architectures" in ADSs that can harm to users [67], and ethnographic researchers in the field who may be necessary to gather empirical social scientific data about algorithmic harms as experienced within communities [96]. The crucial questions to answer here are: How to assemble these diverse forms of expertise and knowledge? How to negotiate access to the code, data, infrastructure, and target populations for these systems? And, how to translate these forms of knowledge into actionable descriptions of concrete impacts?

Expertise is also not limited to professional capacities. Individuals and communities affected by algorithmic systems are often the foremost experts in the potential harms they regularly encounter, as well as the strategies they have developed to minimize or avoid such harms [7]. Indeed, to paraphrase criminal justice reformer Glenn Martin, those closest to the harm are closest to the solution [73]. Without incorporating deeply situated knowledges, the "problematic background assumptions"(p. 787, [52] cited in [91]). that characterize disciplines, and particularly affect the tech industry [74], will continue to produce significant blindspots when it comes to rendering potential harms as measurable impacts. While affected communities' expertise is not a "design-fix" for the potential harms of algorithmic systems, and must be appropriately acknowledged and compensated [100], it is an important component of the form of consensus that resolves the assessor's regress.

Another lesson that can be drawn here is that consensus-building is crucial to both the project of aligning impacts as closely as possible with harms and ensuring robust accountability through the evaluation of algorithmic systems' impacts in ways that mitigate their potential harms. Consensus is both an epistemic commitment and a governance commitment. A regulatory agency cannot stipulate by fiat what should and what should not be included in an AIA.

Regulatory agencies do not have adequate access to the types of grounded research, technical expertise, or entrée to affected communities to ensure that any list of impacts it stipulates is adequate to the potential harms people may experience. Neither can a private company that develops algorithmic systems evaluate the comprehensiveness of its own efforts to evaluate impacts, for three reasons: (1) the possibility of legal endogeneity to reproduce the company's own definitions of compliance as adequate to regulatory requirements; (2) limitations in the types of impacts made measureable by the instrumentation built into algorithmic products; and (3) the absence of an external forum that can mandate changes to the system. Nor can critical academic, journalistic, or other third party audits stand in as impact assessments, despite the fact that such audits can reveal important harms unanticipated by algorithmic systems' designers. Independent critical audits lack access to the design specifications and internal technical description of algorithmic systems that are often protected by intellectual property laws, and they do not in themselves have direct influence over the design and operation of algorithmic systems (even if audits and journalistic investigations have occasionally prompted public outcry, industry responses, and legislative action, and regulatory scrutiny in the recent past).

In terms of epistemic commitments, none of these entities—regulatory agencies, private companies, affected communities, and independent investigators—alone possesses enough insight into the design, operation, and effects of algorithmic systems to be able to evaluate the relationship between their impacts and their actual or potential harms. In terms of governance commitments, none of these entities alone possesses the authority to assess the adequacy of those impacts or demand changes to systems that produce unacceptable impacts. However, these entities, together, form an epistemic community [99] capable of resolving the assessor's regress described above, and could form the basis for meaningful accountability.

5 CONCLUSION

In this paper we have analyzed the history of impact assessment in multiple domains to argue that "impacts" can best be understood as evaluative constructs that emerge from and through (i.e., co-constructed) relationships of accountability. In the context of algorithmic governance, relationships of accountability are crucial for structuring the associations between actors and fora in ways that allow the fora to demand changes to an algorithmic system. But, we argue, these relationships are also crucial for interrogating the ways in which impacts are constructed by a range of relevant actors to make harms to people and communities legible. We point to the necessity of combining diverse forms of expertise, and foregrounding the expertise of communities most likely to be affected by algorithmic systems, so that consensus can become a resource for bringing measurable impacts as close to potential harms as possible. Consensus across forms of expertise and affected communities offers an escape from what we call the "assessor's regress", which occurs when uncertainty arises about the legitimate way to evaluate potential harms as measurable impacts. While AIAs cannot be predicated on the illusion that every harm caused by a given ADS can be foreseen, this paper calls attention to the possible danger that in assessing impacts, AIAs may become an abstract exercise, which does not account for the harms algorithimic systems can engender in practice.

ACKNOWLEDGMENTS

The authors wish to thank our colleagues who took the time to read this work in draft form: Patrick Davidson, Sareeta Amrute, and participants in the Raw Materials Seminar at Data Society Research Institute. We also wish to thank the NSF (awards #1704369 and #1633400) for supporting portions of this work.

REFERENCES

[1] Julius A. Adebayo and others. 2016. *FairML: ToolBox for diagnosing bias in predictive modeling*. Ph.D. Dissertation. Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/108212

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs]* (June 2016). http://arxiv.org/abs/1606.06565 arXiv: 1606.06565.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. *ProPublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] K.A. Bamberger and Deirdre K Mulligan. [n.d.]. PIA requirements and privacy decision-making in US government agencies. In *Privacy Impact Assessment*. Springer, Dordrecht, 225–250.

[5] Karen Barad. 2003. Posthumanist Performativity Toward an Understanding of How Matter Comes to Matter. *Signs: Journal of Women in Culture and Society* 28, 3 (2003).

[6] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal* (2016). https://doi.org/10.2139/ssrn.2477899

[7] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, Medford, MA.

[8] F. Bieker, M. Friedewald, M. Hansen, H. Obersteller, and M. Rost. 2016. A process for data protection impact assessment under the european general data protection regulation. In *Annual Privacy Forum*. Frankfurt, 21–37.

[9] Wiebe E. Bijker. 1995. *Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change*. MIT Press, Cambridge, MA.

[10] Wiebe E. Bijker, Thomas Parke Hughes, and Trevor Pinch (Eds.). 1987. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, Mass.

[11] Wiebe E. Bijker and John Law (Eds.). 1992. *Shaping Technology/Building Society: Studies in Sociotechnical Change*. MIT Press, Cambridge, MA.

[12] Reuben Binns. 2019. On the Apparent Conflict Between Individual and Group Fairness. *arXiv:1912.06883 [cs, stat]* (Dec. 2019). http://arxiv.org/abs/1912.06883 arXiv: 1912.06883.

[13] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121* (2016). https://arxiv.org/abs/1606.06121

[14] Sen. Cory Booker. 2019. Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms. https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms

[15] Mark Bovens. 2007. Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal* 13, 4 (2007), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0386.2007.00378.x.

[16] Geoffrey C. Bowker. 2013. Data flakes: An afterword to "raw data" is an oxymoron. *Raw data" is an oxymoron* (2013), 167–171.

[17] danah boyd. 2014. The Networked Nature of Algorithmic Discrimination. In *Data and Discrimination: Collected Essays*, Seeta Peña Gangadharan, With Virginia Eubanks, and Solon Barocas (Eds.). Open Technology Institute, 6.

[18] Harry Brignull. 2013. Dark Patterns: inside the interfaces designed to trick you. https://www.theverge.com/2013/8/29/4640308/dark-patterns-inside-the-interfaces-designed-to-trick-you

[19] BSR. 2018. *Human Rights Impact Assessment: Facebook in Myanmar*. Technical Report. https://fbnewsroomus.files.wordpress.com/2018/11/bsr-facebook-myanmar-hria_final.pdf

[20] Robert D Bullard. 1999. Dismantling Environmental Racism in the USA. *Local Environment* 4, 1 (1999), 25.

[21] Joy Buolamwini. 2019. Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces. https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced

[22] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, Vol. 18.

[23] Robert W. Burchell, David Listokin, and William R. Dolphin. 1985. *The New Practitioner's Guide to Fiscal Impact Analysis*. Center for Urban Policy Research, New Brunswick, NJ.

[24] Robert W. Burchell, David Listokin, William R. Dolphin, Lawrence Q. Newton, and Susan J. Foxley. 1994. *Development Impact Assessment Handbook*. Urban Land Institute, Washington, DC.

[25] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. 134–148. http://proceedings.mlr.press/v81/chouldechova18a.html

[26] Roger Clarke. 2009. Privacy impact assessment: Its origins and development. *Computer law & security review* 25, 2 (2009), 123–135.

[27] Yvette D. Clarke. 2019. H.R.2231 - 116th Congress (2019-2020): Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231 Archive Location: 2019/2020.

[28] Luke W. Cole. 1992. Remedies for Environmental Racism: A View from the Field. *Michigan Law Review* 90, 7 (June 1992), 1991. https://doi.org/10.2307/1289740

[29] Harry Collins. 1985. *Changing Order: Replication and Induction in Scientific Practice*. Sage, London.

[30] U.S. Congress. 1969. The National Environmental Policy Act of 1969. https://www.fws.gov/r9esnepa/RelatedLegislativeAuthorities/nepa1969.PDF

[31] Noel Corriveau. 2018. The Government of Canada's Algorithmic Impact Assessment: Take Two. https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f Library Catalog: medium.com.

[32] Tim Dare and Eileen Gambrill. 2016. Ethical analysis: Predictive risk models at call screening for Allegheny County. *Unpublished report* (2016).

[33] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (Jan. 2016), 56–62. https://doi.org/10.1145/2844110

[34] Dylan Doyle-Burke and Jessie Smith. [n.d.]. IBM, Microsoft, and Amazon Disavow Facial Recognition Technology: What Do You Need to Know? with Deb Raji. https://radicalai.podbean.com/e/ibm-microsoft-and-amazon-disavow-facial-recognition-technology-what-do-you-need-to-know-with-deb-raji/

[35] Lauren B. Edelman and Shauhin A. Talesh. 2011. To Comply or Not to Comply – That Isn't the Question: How Organizations Construct the Meaning of Compliance. In *Explaining Compliance*. Edward Elgar Publishing. https://doi.org/10.4337/9780857938732.00011

[36] Wendy Nelson Espeland and Berit Irene Vannebo. 2007. Accountability, Quantification, and Law. *Annual Review of Law and Social Science* 3, 1 (Dec. 2007), 21–43. https://doi.org/10.1146/annurev.lawsocsci.2.081805.105908

[37] Ana Maria Esteves, Daniel Franks, and Frank Vanclay. 2012. Social impact assessment: the state of the art. *Impact Assessment and Project Appraisal* 30, 1 (March 2012), 34–42. https://doi.org/10.1080/14615517.2012.660356

[38] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[39] 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*. ACM, 259–268. http://dl.acm.org/citation.cfm?id=2783311

[40] Panel for the Future of Science and Technology. 2019. *A governance framework for algorithmic accountability and transparency.* Technical Report. European Parliamentary Research Service, LU. https://data.europa.eu/doi/10.2861/59990

[41] Lisa Forman and Gillian MacNaughton. 2016. Lessons learned: a framework methodology for human rights impact assessment of intellectual property protections in trade agreements. *Impact Assessment and Project Appraisal* 34, 1 (Jan. 2016), 55–71. https://doi.org/10.1080/14615517.2016.1140995

[42] William R Freudenburg. [n.d.]. Social Impact Assessment. ([n. d.]), 28.

[43] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]* (Sept. 2016). http://arxiv.org/abs/1609.07236 arXiv: 1609.07236.

[44] Oscar H. Gandy. 2010. Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems. *Ethics and Information Technology* 12, 1 (March 2010), 29–42. https://doi.org/10.1007/s10676-009-9198-6

[45] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2018). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[46] C. A. E. Goodhart. 1984. *Monetary Theory and Practice.* Macmillan Education UK, London. https://doi.org/10.1007/978-1-349-17295-5

[47] Ben Green and Salomé Viljoen. 2020. Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona, ES, 13.

[48] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7 (July 2013), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

[49] Sandra Harding. 1992. After the Neutrality Ideal: Science, Politics, and "Strong Objectivity". *Social Research* 59, 3 (1992), 22.

[50] Libby Hogan and Michael Safi. 2018. Revealed: Facebook hate speech exploded in Myanmar during Rohingya crisis. *The Guardian* (April 2018). https://www.theguardian.com/world/2018/apr/03/revealed-facebook-hate-speech-exploded-in-myanmar-during-rohingya-crisis

[51] IAIA. 2009. What is Impact Assessment.

[52] Kristen Intemann. 2010. 25 Years of Feminist Empiricism and Standpoint Theory: Where Are We Now? *Hypatia* 25, 4 (2010), 778–796. https://doi.org/10.1111/j.1527-2001.2010.01138.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1527-2001.2010.01138.x.

[53] Sheila Jasanoff. 2004. *States of knowledge: the co-production of science and the social order.* Routledge.

[54] Stephen Jay, Carys Jones, Paul Slinn, and Christopher Wood. 2007. Environmental impact assessment: Retrospect and prospect. *Environmental Impact Assessment Review* 27, 4 (May 2007), 287–300. https://doi.org/10.1016/j.eiar.2006.12.001

[55] Khari Johnson. 2020. Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI. *VentureBeat* (Sept. 2020). https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/

[56] Margot E. Kaminski and Gianclaudio Malgieri. 2019. Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3456224

[57] Michael Karlin. [n.d.]. Author biography. https://policyoptions.irpp.org/authors/michael-karlin/

[58] Michael Karlin. 2018. The Government of Canada's Algorithmic Impact Assessment: Take Two. https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f

[59] Sonia K Katyal. 2019. Private Accountability in the Age of Artificial Intelligence. *UCLA Law R.* 66 (2019), 54.

[60] Deanna Kemp and Frank Vanclay. 2013. Human rights and impact assessment: clarifying the connections in practice. *Impact Assessment and Project Appraisal* 31, 2 (June 2013), 86–96. https://doi.org/10.1080/14615517.2013.782978

[61] Zenia Kotval and John Mullin. 2006. *Fiscal Impact Analysis: Methods, Cases, and Intellectual Debate.* Technical Report. Lincoln Institute of Land Policy.

[62] Mark Latonero. 2020. Can Facebook's Oversight Board Win People's Trust? *Harvard Business Review* (Jan. 2020), 4. https://hbr.org/2020/01/can-facebooks-oversight-board-win-peoples-trust

[63] Bruno Latour. 2005. *Reassembling the social: an introduction to actor-network-theory.* Oxford University Press, Oxford ; New York. OCLC: ocm58054359.

[64] David P. Lawrence. 2000. Planning theories and environmental impact assessment. *Environmental Impact Assessment Review* 20, 6 (Dec. 2000), 607–625. https://doi.org/10.1016/S0195-9255(00)00036-6

[65] F. Larry Leistritz. 1994. Economic and Fiscal Impact Assessment. *Impact Assessment* 12, 3 (1994), 305–317. https://doi.org/10.1080/07349165.1994.9725868

[66] Mathieu Lemay. 2019. Understanding Canada's Algorithmic Impact Assessment Tool. https://towardsdatascience.com/understanding-canadas-algorithmic-impact-assessment-tool-cd0d3c8cafab

[67] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R. Sunstein, and Ralph Hertwig. 2020. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour* (June 2020). https://doi.org/10.1038/s41562-020-0889-7

[68] Michael Lynch. 2016. Social Constructivism in Science and Technology Studies. *Human Studies* 39, 1 (March 2016), 101–112. https://doi.org/10.1007/s10746-016-9385-5

[69] Donald MacKenzie. 1989. From Kwajalein to Armageddon? Testing and the Social Construction of Missle Accuracy. In *The Uses of Experiment: Studies in the Natural Sciences*, David Gooding, Trevor Pinch, and Simon Schaffer (Eds.). Cambridge University Press, New York, 409–436.

[70] Donald A. MacKenzie and Judy Wajcman (Eds.). 1985. *The Social shaping of technology: how the refrigerator got its hum.* Open University Press, Milton Keynes ; Philadelphia.

[71] Koray Mancuhan and Chris Clifton. 2014. Combating discrimination using Bayesian networks. *Artificial Intelligence and Law* 22, 2 (June 2014), 211–238. https://doi.org/10.1007/s10506-014-9156-4

[72] Alessandro Mantelero. 2018. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review* 34, 4 (Aug. 2018), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

[73] Glenn E. Martin. 2017. Those closest to the problem are closest to the solution. https://theappeal.org/those-closest-to-the-problem-are-closest-to-the-solution-555e04317b79/

[74] Jacob Metcalf, Emanuel Moss, and danah boyd. 2019. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research* 86, 2 (2019), 449–476.

[75] Tom Miles. 2018. U.N. investigators cite Facebook role in Myanmar crisis. *Reuters* (2018), 10.

[76] 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. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (2019), 220–229. https://doi.org/10.1145/3287560.3287596 arXiv: 1810.03993.

[77] Peter Montague. 2004. Reducing the harms associated with risk assessments. *Environmental Impact Assessment Review* 24, 7-8 (Oct. 2004), 733–748. https://doi.org/10.1016/j.eiar.2004.06.004

[78] Richard K. Morgan. 2012. Environmental impact assessment: the state of the art. *Impact Assessment and Project Appraisal* 30, 1 (March 2012), 5–14. https://doi.org/10.1080/14615517.2012.661557

[79] Peter Morris and Riki Therivel. 2001. *Methods of environmental impact assessment.* Spon Press, London; New York. http://site.ebrary.com/id/5001176 OCLC: 50016773.

[80] Emanuel Moss and Jacob Metcalf. 2020. High Tech, High Risk: Tech Ethics Lessons for the COVID-19 Pandemic Response. *Patterns* 1, 7 (2020).

[81] Emanuel Moss, Elizabeth Anne Watkins, Jacob Metcalf, and Madeleine Clare Elish. 2020. Governing with Algorithmic Impact Assessments: Six Observations. *SSRN Electronic Journal* (2020).

[82] Government of Canada. 2002. E-Government Act of 2002. https://www.govinfo.gov/content/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf

[83] Government of Canada. 2020. aia-eia-js/survey-enfr.json at master · canada-ca/aia-eia-js · GitHub. https://github.com/canada-ca/aia-eia-js/blob/master/src/survey-enfr.json

[84] Government of Canada. 2020. Algorithmic Impact Assessment - Évaluation de l'Incidence Algorithmique. https://canada-ca.github.io/aia-eia-js/

[85] Government of Canada. 2020. Canadian Treasury Board Directive on the Use of Machine Learning for Decision-Making. https://docs.google.com/document/d/1LdciG-UYeokx3U7ZzRng3u4T3IHrBXXk9JddjjueQok/edit

[86] Council of Europe. 2016. General Data Protection Regulation (GDPR) – Official Legal Text. https://gdpr-info.eu/

[87] Judith Petts. 1999. *Handbook of Environmental Impact Assessment Volume 2: Impact and Limitations.* Vol. 2. Blackwell Science, Oxford.

[88] Theodore M. Porter. 1995. *Trust in numbers: the pursuit of objectivity in science and public life.* Princeton University Press, Princeton, N.J.

[89] Michael Power. 1994. *The Audit Explosion.* Technical Report. Demos, London: Demos. https://www.demos.co.uk/files/theauditexplosion.pdf

[90] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19).* Association for Computing Machinery, New York, NY, USA, 429–435. https://doi.org/10.1145/3306618.3314244

[91] 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 *Conference on Fairness, Accountability, and Transparency (FAT* '20),*. Barcelona, ES, 12.

[92] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.* Technical Report. AI Now. 22 pages.

[93] Andrea Romei, Salvatore Ruggieri, and Franco Turini. 2013. Discrimination discovery in scientific project evaluation: A case study. *Expert Systems with Applications* 40, 15 (Nov. 2013), 6064–6079. https://doi.org/10.1016/j.eswa.2013.05.016

[94] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data* 4, 2 (May 2010), 1–40. https://doi.org/10.1145/1754428.1754432

[95] Johan Schot and Arie Rip. 1997. The Past and Future of Constructive Technology Assessment. *Technological Forecasting and Social Change* 54, 2-3 (1997), 251–268. https://doi.org/10.1016/S0040-1625(96)00180-1

[96] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (Dec. 2017), 2053951717738104. https://doi.org/10.1177/2053951717738104 Publisher: SAGE Publications Ltd.

[97] Andrew D. Selbst. 2018. Disparate Impact in Big Data Policing. *Georgia Law Review* 52 (2018), 109. https://doi.org/10.2139/ssrn.2819182

[98] 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 - FAT* '19*. ACM Press, Atlanta, GA, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[99] Steven Shapin, Simon Schaffer, and Thomas Hobbes. 1985. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life: including a translation of Thomas Hobbes, Dialogus physicus de natura aeris by Simon Schaffer*. Princeton University Press, Princeton, N.J.

[100] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a Design Fix for Machine Learning. In *Proceedings of the 37th International Conference on Machine Learning*. Vienna, Austria, 7.

[101] Susan Leigh Star. 1989. The Structure of Ill-Structured Solutions: Boundary Objects and Heterogenous Distributed Problem Solving. In *Distributed artificial intelligence*, L. Gasser and M. Huhns (Eds.). Pitman, London.

[102] Alexandra Stevenson. 2018. Facebook Admits It Was Used to Incite Violence in Myanmar (Published 2018). *The New York Times* (Nov. 2018). https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html

[103] D. Tancock, S. Pearson, and A. Charlesworth. 2010. *The emergence of privacy impact assessments*. Technical Report. Hewlett Packard. http://www.hpl.hp.com/techreports/2010/HPL-2010-63.pdf

[104] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. 2017. *Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation*. Technical Report. 60 pages. https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf

[105] Kush R. Varshney. 2018. Introducing AI Fairness 360, A Step Towards Trusted AI - IBM Research. https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/ Library Catalog: www.ibm.com Section: AI.

[106] 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, Barcelona Spain, 1–18. https://doi.org/10.1145/3351095.3372833

[107] Daniel C Wigley and Kristin S Shrader-Frechette. 1996. Environmental Racism and Biased Methods of Risk Assessment. *RISK* 7 (1996), 55–88.

[108] Caitlin Zaloom. 2004. The Productive Life of Risk. *Cultural Anthropology* 19, 3 (Aug. 2004), 365–391. https://doi.org/10.1525/can.2004.19.3.365

[109] Tal Z. Zarsky. 2017. An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics. *ISJLP* 14 (2017), 11.