Governing Algorithmic Systems with Impact Assessments: Six Observations

Elizabeth Anne Watkins
ewatkins@datasociety.net
Data & Society Research Institute
New York, NY
Princeton Center for Information Technology Policy, Princeton, NJ

Emanuel Moss
emanuel@datasociety.net
Data & Society Research Institute
New York, NY
CUNY Graduate Center
New York, NY

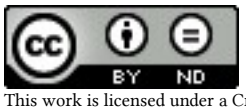
Jacob Metcalf
jake.metcalf@datasociety.net
Data & Society Research Institute
New York, NY

Ranjit Singh
ranjit@datasociety.net
Data & Society Research Institute
New York, NY

Madeleine Clare Elish*
mcelish@google.com
Google Research
Mountain View, CA

Abstract

Algorithmic decision-making and decision-support systems (ADS) are gaining influence over how society distributes resources, administers justice, and provides access to opportunities. Yet collectively we do not adequately study how these systems affect people or document the actual or potential harms resulting from their integration with important social functions. This is a significant challenge for computational justice efforts of measuring and governing AI systems. Impact assessments are often used as instruments to create accountability relationships and grant some measure of agency and voice to communities affected by projects with environmental, financial, and human rights ramifications. Applying these tools—through Algorithmic Impact Assessments (AIA)—is a plausible way to establish accountability relationships for ADSs. At the same time, what an AIA would entail remains under-specified; they raise as many questions as they answer. Choices about the methods, scope, and purpose of AIAs structure the conditions of possibility for AI governance. In this paper, we present our research on the history of impact assessments across diverse domains, through a sociotechnical lens, to present six observations on how they co-constitute accountability. Decisions about what type of effects count as an impact; when impacts are assessed; whose interests are considered; who is invited to participate; who conducts the assessment; how assessments are made publicly available, and what the outputs of the assessment might be; all shape the forms of accountability that AIAs engender. Because AIAs are still an incipient governance strategy, approaching them as social constructions that do not require a single or universal approach offers a chance to produce interventions that emerge from careful deliberation.



AIES '21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05. https://doi.org/10.1145/3461702.3462580

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

KEYWORDS
algorithmic impact assessment; impact; harm; accountability; governance

ACM Reference Format:
Elizabeth Anne Watkins, Emanuel Moss, Jacob Metcalf, Ranjit Singh, and Madeleine Clare Elish. 2021. Governing Algorithmic Systems with Impact Assessments: Six Observations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21) May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA. https://doi.org/10.1145/3461702.3462580

Introduction

From government policy makers to company board rooms, the idea of implementing "algorithmic impact assessments" (AIAs) as a form of algorithmic accountability is gaining momentum. These assessments are seen as potentially useful for anticipating, avoiding, and mitigating the negative consequences of algorithmic decision systems (ADS[1]) (Selbst 2017, Reisman et al. 2018). Already, the EU has stipulated through its GDPR legislation, in addition to its other recent legislative efforts (Johnson 2020), that in the interest of user rights, companies must provide privacy impact assessments upon request (Kaminski and Malgieri 2019). The Canadian government now requires a checklist-style version of algorithmic impact assessments for its agencies that use algorithms (Karlin and Corriveau 2018, Government of Canada 2019). Companies like Facebook and Google are commissioning human rights impact assessments to identify harms

[1] For the purposes of this research, we will refer to algorithmic decision-making and decision-support systems collectively as ADSs.
* Research and writing was conducted primarily when the author was a Program Director at Data & Society Research Institute.

of their platforms and products (Warofka 2018, Allison-Hope et al. 2019). The Algorithmic Accountability Act, proposed in the US Congress in 2019, would require companies with large user-bases to conduct impact assessments of their ADSs that affect certain sensitive domains of people's lives (H.R. 2231 2019).

The term "algorithmic impact assessment" (AIA) has been used as an umbrella term, referring to a range of processes and documentation. It emerges within the context of an expanding toolbox of potential accountability processes, including algorithmic audits, datasheets, "nutrition" labels, and model cards (Raji 2020, Gebru et al. 2018, Holland et al. 2018, Mitchell et al. 2019). The general idea of an AIA is to document the development and impact of an ADS, providing a point of leverage for mitigating potential harms to individuals and communities, particularly vulnerable individuals and communities.[2] It certainly is a compelling intervention, but it leaves more questions than answers. What constitutes an assessment? An impact? An algorithm? An ADS? Who gets to decide? Should algorithms used by private companies be subject to the same forms of accountability as those used by public institutions? What forms of accountability are at stake?

Existing proposals for AIAs and related governance practices answer each of these questions differently. This is to be expected. There is not as of yet a clear coalescence of institutional, intellectual, regulatory, and judicial power around any particular vision of what an AIA *is*. This heterogeneity provides an important opportunity to critically shape the purpose and methods of AIAs in the future.

ADSs, which encompass machine learning and AI techniques, present unique and substantial challenges when it comes to assessing their impact on society. These include, but are not limited to, how these systems are built, how they relate to the data used to train and retrain them, and the power relationships between agencies and industries that operate ADSs, the complex role played by 3rd-party vendors, and how "users" and "the public" are constituted (Cath 2018, Koene et al. 2019, Veale and Brass 2019, Mulligan and Bamberger 2019). The existing body of research on how to audit, investigate, and understand undesirable and unexpected behaviors of such systems is currently growing, and is much needed. Moreover, there is a lack of empirical evidence and research to support how—or whether—AIA will become an effective, or even a desirable, governance mechanism.[3] A robust approach to AIAs will couple these ongoing efforts with considerations around the range of social, technological, and legal entities implicated in its process.

If the goal is to develop new and stronger mechanisms of accountability for the cascading effects of ADSs, impact assessments offer many opportunities. Rather than relying on slow-moving legislatures to outline exactly what ADSs can and cannot do, AIAs—whether mandated or directly administered by a responsible government agency—can set standards for evaluating the performance of such systems. It provides a means of accountability that tracks alongside the shift in power as it moves from lawmakers to agency personnel to those who perform impact assessments (Shapiro 1965, DeLong 1979, West 2005). Impact assessments provide a basis for rational decision-making between competing alternatives in the design of a development project, where tradeoffs between potential upside benefits and downside impacts must be made (Steinemann 2001).

At the same time, the efficacy of impact assessments has been critiqued in the context of ADSs, as well as in other domains, including fiscal, privacy and environmental impact assessments (Mauer 2007, Kaminksi and Malgieri 2019).[4] These critiques have focused on the role that impact assessments play in creating governance mechanisms that abet superficial self-regulation or the mere veneer of accountability (Waldman 2020, Mourey and Waldman 2020).

Can AIAs be effective governance and accountability mechanisms for ADSs, and if so, how? Clearly, the motivation of impact assessment is to identify, measure, balance, minimize, and mitigate harms; harms to people are at the heart of impact assessment. Impact assessment is necessarily a form of institutional governance, prone to rendering questions of ethics and justice as compliance exercises. At the heart of potential AIA practices is a tension between the need to identify often invisible and disaggregated harms to protect populations, and the need to render these harms in formal language legible to institutions to act upon. We argue that the challenge at hand for AIAs is how to practically achieve legitimacy—through robust methods, cross-disciplinary expertise, and inclusive participatory structures—in bridging this tension.

As we argue below, and in greater depth elsewhere (Moss et al. 2021), the accountability relationship between an *actor* who builds and designs systems and a *forum* capable of demanding changes is central to operationalizing an impact assessment

[2] The Ada Lovelace Institute and Datakind UK have pointed out that algorithm auditing (i.e. how does the system function and is it accurately described?) despite being both more robustly fleshed-out (especially bias auditing) and having a narrower purview is often conflated with impact assessments (Ada Lovelace Institute and DataKindUK, 2020).

[3] Following the proposal of the Algorithmic Accountability Act, several organizations voiced their concerns about the legislative proposal. See, New, Joshua. 2019. "How to Fix the Algorithmic Accountability Act." Center for Data Innovation: https://www.datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act/; Barbanel, Jerry. 2019. "A look at the the proposed Algorithmic Accountability Act of 2019." IAPP.org: https://iapp.org/news/a/a-look-at-the-proposed-algorithmic-accountability-act-of-2019/; Selbst, Andrew, Madeleine Clare Elish and Mark Latonero. "Accountable Algorithmic Futures." *Points*. https://points.datasociety.net/building-empirical-research-into-the-future-of-algorithmic-accountability-act-d230183bb826.

[4] See Taylor 1983 for a canonical study of how the NEPA environmental impact assessment process facilitates development projects despite foundational intentions to balance competing interests of environmentalists and federal agencies. See also Goldman, Michael. 2005 (Imperial Nature: The World Bank and Struggles for Social Justice in the Age of Globalization. Yale Agrarian Studies Series. New Haven, CT: Yale University Press) for a study of the World Bank's use of an environmental impact assessment process to depoliticize international development and "greenwash" exploitative economic development projects in the developing world.

regime. How will harms be rendered as impacts, which actor will be responsible to report those impacts to which forum, who will have access to the forum, when will those impacts be assessed, are all matters of how (and whether) an AIA regime is able to maintain legitimacy by adequately co-constructing aspects of institutional compliance to the very real harms that people experience from ADSs. In other words, the *reality* of impacts as an evaluative construct relies on the ability of institutions to construct impacts *well*. Fortunately, as we will show in later sections, there are resources available for understanding how the constructedness of a shared object is entangled with the ethical obligation of assembling it.

As a contribution to the growing area of inquiry and action on how AIAs can be effectively assembled, we draw on our backgrounds in the social sciences and our experience studying and analyzing the consequences of ADSs to think through the recent history of impact assessment and identify lessons that might be learned for AIAs. We identify how an AIA process might reasonably reduce harms to individuals and groups and minimize disruptive impacts, while still producing useful and beneficial ADSs. We purposefully avoid making prescriptions here about the best way to assemble AIA processes, although we assume that the government as well as the private sector and civil society, will have key roles to play. To this end, we offer six observations on AIAs as instruments for assembling accountability:

- What constitutes an impact is non-obvious.
- Different types of impact come into focus depending on when an assessment occurs.
- Public participation in an assessment process is not synonymous with accountability to the public.
- Impact assessments structure how institutions operate and interact.
- Assessing impacts does not necessarily mean addressing harms.
- Impact assessments ask us how the world might be otherwise.

Methods

We conducted historical research on impact assessment regimes through a sociotechnical lens for this paper. We surveyed available materials in governance and legislation, such as bills, federal agency guidelines and documents, impact statements, and EU documents. We also surveyed available critiques on these structures from legal and sociological disciplines. We then surveyed contemporary proposals around assessments for ADSs, including methods for internal and independent audits and end-to-end frameworks. Throughout this research, we focused on how the constructedness of evaluative objects like impact assessments pragmatically relate to the material harms they were intended to assess. This focus led us to analyze how organizational commensuration practices shape accountability relationships between institutions, technical systems, and social structures.

Constituting Impact

In this section, we describe several common features—that we have come to call constitutive components—of impact assessments across diverse domains. In identifying these components, we found that how these components are assembled inevitably constitutes the efficacy of impact assessment as a tool for structuring accountability relationships.

As evaluative constructs, impact assessments enable actors to identify and address harms. Harms, thus, are not identical to impacts; rather, impact assessment practices constitute "impacts" as proxies for actual or potential harms. An impact assessment regime configures certain systems as capable of causing impacts. It stabilizes abstract concepts (like rights, environmental resources, or privacy) as tangible entities capable of being impacted in concrete and measurable ways. Impact assessments, thus, materially instantiate social and political priorities about what is worthy of assessment and what types and levels of harms are tolerable. Therein lies a risk: when "impacts" are constructed as institutionally legible proxies for harms, what is ultimately documented and addressed as "impact" may become distant from the actual (or potential) harms experienced by people. In other words, how impacts get constructed as representations of harms has significant ethical stakes.

Furthermore, impacts are *co-constructed* (Bijker et al. 1987, Bijker 1995, Jasanoff 2004, Latour 2005, Lynch 2016) with the accountability relationships that structure responsibility for harms and their amelioration. The *what* of an impact is inextricably tied to *who* is responsible for measuring and mitigating the impact. According to the definition of public accountability that commonly circulates in algorithmic accountability literatures, accountability within and between institutions is contingent on a five-part ecosystem: 1) an actor submitting a technical account of the impact of a system; 2) a forum that evaluates that account and can propose suitable changes to the system; 3) the structured relationship between the actor and the forum; 4) the criteria for assessing impact, and 5) the consequences arising from these accounts (Bovens 2007, Wieringa 2020). Current and proposed forms of AIAs vary widely in terms of who the accountable actors and fora should be, what the relationship between them should be, the criteria for assessment, and the relationship between assessments and the consequences arising therefrom.

Alongside the actors and the forum to establish accountability, impact assessments share a number of other constitutive components outlined in **Table 1** (Moss et al. 2021). While details vary, these components appear consistently across domains. We analyzed impact assessment processes across five domains: Human Rights Impact Assessment (HRIA), Data Protection Impact Assessment (DPIA), Fiscal Impact Assessment (FIA), Privacy Impact Assessment (PIA), and Environmental Impact Assess-

ment (EIA) to identify these ten components. Every impact regime is characterized by an assembly of these components. AIAs, thus, require a stable and common response to each of them to be effective governance mechanisms.

Components	**Implications for governing with impact assessments**
Sources of Legitimacy	IAs* can only be effective in establishing accountability relationships when they are legitimized either through legislation or within a set of norms that are officially recognized and publicly valued. Without a *source of legitimacy*, IAs may fail to provide a forum the power to impute responsibility to actors.
Actors and Forum	IAs are rooted in establishing an accountability relationship between *actors* that design, deploy, and operate a system and a *forum* that can allocate responsibility for potential consequences of such systems and demand changes in their design, deployment, and operation.
Catalyzing Event	*Catalyzing events* are triggers for conducting IAs. These can be mandated by law or solicited voluntarily at any stage of a system's development life cycle. Such events can also manifest through on-the-ground harms from a system's operation experienced at a scale that cannot be ignored.
Time Frame	Once an IA is triggered, *time frame* is the period often mandated through law or mutual agreement between actors and the forum within which an IA must be conducted. Most IAs are performed *ex ante*, before developing a system, but they can also be done *ex post* as an investigation of what went wrong.
Public Access	The broader the *public access* to an IA's processes and documentation, the stronger is its potential to enact accountability. Public access is essential to achieving transparency in the accountability relationship between actors and the forum.
Public Participation	*Public participation* creates conditions for solicitation of feedback from the broadest possible set of stakeholders in a system. Such participation is a resource to expand the list of impacts assessed or to shape the design of a system. Who constitutes this public and how their participation is solicited are critical to the success of an IA.
Method	*Methods* are standardized techniques of evaluating and foreseeing how a system would operate in the real-world. For example, public consultation is a common method for IAs. Most IAs have a roster of well-developed techniques that can be applied to foresee the potential consequences of deploying a system as impacts.
Assessors	An IA is conducted by *assessors*. The independence of assessors from the actor as well as the forum is crucial to how an IA identifies impacts, how those impacts relate to tangible harms, and how it acts as an accountability mechanism that avoids, minimizes, or mitigates such harms.
Impacts	*Impacts* are abstract and evaluative constructs that can act as proxies for harms produced through the deployment of a system in the real-world. They enable the forum to identify and ameliorate potential harms, stipulate conditions for system operation, and thus, hold the actors accountable.
Harms and Redress	*Harms* are lived experiences of the adverse consequences of a system's deployment and operation in the real-world. Some of these harms can be anticipated through IAs, others cannot be foreseen. *Redress* procedures must be developed to complement any harms identified through IA processes to secure justice.
*IAs is an acronym for Impact Assessments	

Table 1: The Ten Constitutive Components of Impact Assessment (IA) Processes

The question at hand is which configuration of these components for AIAs will effectively bridge the need to understand the actual harms experienced by people and the imperatives of institutions that must act to mitigate those harms. The harms of ADSs pose distinct challenges (as would any new domain for impact assessment). They are often dispersed, aggregate, and not easily anticipated: (1) they may happen at a distance and to a widely spread population (Keyes 2018, Selbst 2017, Hoffman 2020), (2) be altered by the scale of the system or datasets (Hanna and Park 2020), (3) be introduced through the manipulation of algorithmic parameters for experimental purposes, such as in social media newsfeeds (Tufekci 2015), (4) be brought about by the collision of algorithmic decision-making with unanticipated contextual social or technical parameters (Elish and Watkins 2020), (5) be implicated in complex and new frames of moral and legal responsibility (Datta et al. 2018, Elish 2019, Keddell 2019), or (6) be the result of multiple layers of individually-managed ADSs (O'Brien 2020). Any legitimate algorithmic governance mechanism would need to address the breadth of possible harms that might arise through these means but doing so exceeds the

boundaries of how impact assessment regimes have typically grappled with harms.

We argue that recognition of the "constructedness" of impacts can provide a generative pathway to bring them closer to representing material harms. Recognizing the socially constructed character of impact assessment early in the development of a new assessment process offers a chance to produce interventions that emerge from careful, politically engaged deliberation, which must include input from a diverse community of varying expertise. However, this "constructedness" also manifests its own challenges. In the following section, we lay out six observations on these challenges and the ethical concerns they raise in practice.

Six Observations

The constitutive components laid out in the previous section showcase the making of impacts as evaluative constructs in organizing the relations between actors and the forum. In operationalizing AIAs as mechanisms of governance, the way in which these components are accounted for and assembled shapes the nature of the accountability relations they produce. In this section, we make six observations on the ways in which practices of governance that rely on these constitutive components may fail. While some of these observations seem to follow from an intuitive understanding of the common features of the history of impact assessments across domains, they all are deeply consequential for empirically describing the ethics of deploying AIAs as a mechanism for governing ADSs. In using 'empirically describing', we draw on the work of Michael Lynch who coined the term 'ethigraphy' to argue for a descriptive focus on ethical decisions as they emerge in practice. 'Ethigraphy' is

> a kind of empirical ethics that examines how technological innovations provide conditions for *ad hoc* pursuit of political and ethical closure. Unlike the promise of classic ethics, the aim is not to repair *ad hoc* decisions with actions grounded in moral principles; instead, it is to investigate the circumscribed and circumstantial way moral agents handle novel conflicts and reconstitute natural and social orders. [...] 'Ethics' is not an isolated specialty, or a domain of pure normative principles remote from the give and take of historical conflict (Lynch 2001: 3).

In making these observations, we do not mean to imply that the challenges they outline are traps that can be avoided in the pursuit of operationalizing AIA. The path to AIA does not go around them, rather it goes through them. As actors and the forum wrestle with the challenges posed by these observations, they produce ethics *as* and *in* practices (Ziewitz 2019) of governing ADSs through AIAs.

What Constitutes an Impact is Non-Obvious

There is no pre-existing or universal definition of an "impact" that can be applied in the context of an impact assessment because there is a central confounding question around delineating the boundaries of the impact to be assessed. "Impact" invokes a causal relationship: an action taken by an organization (or a system operated by such an organization) *causes* an *effect* in the world. *Impacts* are means to account for how organizations change an aspect of the world by making it otherwise. However, it is difficult to delineate such clear causal relationships for most phenomena of interest that need to be measured through impacts. This inevitably raises the question: what can be identified as an impact resulting from one particular cause, and how can that cause be properly identified as having stemmed from a system that an organization has control over?

These questions of delineating boundaries of impacts are most often directed by the contextual domain of the specific type of impact assessment. The impact of an undertaking is assessed by bounding the scope of the assessment itself *to* a particular right, domain, or resource. For example, PIAs examine impacts *to* privacy; HRIAs examine impacts *to* human rights; and EIAs assess impacts *to* the environment.

Delineating boundaries of the impacts of ADSs is difficult in comparison. The domain of operation of an ADS can become expansive over time and thus the domain of any given AIA could be similarly broad. One need to look no further than credit scores to sense just how expansive the set of impacts from an ADS may become. Despite being primarily an algorithmically-generated metric engineered for the purpose of reselling tranches of debt between financial institutions that is only loosely indexed to individuals' financial trustworthiness, credit scores have exhibited considerable function creep. In a process called "off-label use," scores are now being used for purposes far from credit worthiness, including to assess applications for housing and calculate insurance premiums (Rona-Tas 2017). What types of impacts can be reasonably attributed to ADSs, and not to other causes? In case an effect is determined to have multiple causes, how can an assessment attribute a reasonable degree of responsibility to those implementing an ADS? Impact may arise from the data used to train the model, from the algorithmic techniques and design specifications employed in the model, or from the context in which it is applied in the real world. Importantly, the components of an ADS may be assembled from many different sources of data, using many different open and proprietary code bases, and be used in manners quite tangential to their original purposes. Many different parts of a company and/or users of a system may be implicated by different components of an ADS. In thinking through AIAs, it is crucial to answer questions about: (1) what counts as an impact; and (2) how those impacts might be measured and used for any sort of rational evaluation.

Since ADSs are complex and multi-causal, defining what counts as an impact is necessarily a project of prioritization, and

as such is a site of contestation shaped by social, economic, and political power. What constitutes an impact is not only non-obvious but should also be understood as an essentially ethical choice, in which one set of concerns, events, or communities is prioritized over another.

There are no neutral endpoints in the power-laden process of identifying, measuring, formalizing, and accounting for impacts. Which impacts get assessed is ultimately the result of decisions on the assess-ability of a given impact. For instance, measuring the impact of a proposed development project to a city's tax base is easier to quantify and evaluate compared to the project's impact on a neighborhood's feeling of cohesion and community. The list of impacts considered assess-able will remain incomplete, and thus assessments will always be partial.

Different Types of Impact Comes into Focus Depending on When an Assessment Occurs

Critical to unpacking *what* is an "impact" is attending to *when* that impact is assessed. The *time* when an impact assessment is conducted shapes how exhaustive the list of possible impacts of a system can be. While it is impossible to foresee all possible impacts of a system, the practice of anticipating them or accounting for their causes depends on whether an impact is assessed *ex ante* (before deployment) or *ex post* (after deployment).

When impacts are assessed *ex ante*, the assessment is an exercise in predicting the risks and foreseeable consequences of a proposed system. Generally speaking, *ex ante* assessments are based on existing information like prior use cases, empirical measurements of the behavior of the system in testing environments, or narrative records of how the system was designed and iteratively developed.[5] Environmental impact reports, data protection impact assessments, and fiscal impact assessments are based on *ex ante* assessments. In contrast, when impacts are assessed *ex post*, the assessment is an exercise in determining causality of the consequences of an already deployed system or speculating alternative outcomes in an imagined world where the system was not deployed. It is based on a record of information that is primarily gained by following a system's deployment. Generally speaking, this information might include field observations, interviews with stakeholders, or measurements of outcomes in the real world. Examples of impact assessments that use *ex post* assessments are supply chain assessments and human rights impact assessments.

On one hand, proponents of *ex ante* methods often argue that this approach creates invaluable opportunities to assess a project and accordingly modify design prior to its release. In the case of environmental impact assessments, for example, the public debates that occur *before* a development can begin are critical spaces to voice dissent. On the other hand, proponents of *ex post* approaches often argue that it is the impacts that we are least equipped to predict that are the ones that are likely to be most important to observe and assess. These impacts are likely only be knowable *post facto*, when a system has been deployed and integrated in specific social contexts. When and how an impact is assessed not only affects the types of impacts that can be assessed, but also the kinds of processes that need to be established for an assessment to produce organizational accountability. *Ex ante* assessments ask what the anticipated impact of decisions might likely be, while *ex post* assessments ask what would have happened had a different choice been made, and, by implication, is made going forward.

These two forms of assessments rest on differing theories of change, meet different organizational demands, and posit different relationships between cause and effect. They differ in how they view, frame, and describe choices and impact. While *ex post* analyses imagine how an agency or business might intervene in an ongoing process, *ex ante* analyses ask assessors to imagine the potential rewards or risks at stake and must bracket away the difficulties of anticipating the outcomes in the real world (Bailey et al. 2002). Although the approaches can be complementary over the life cycle of a system, assessments are temporally bounded, and there are tradeoffs involved in choosing one approach over the other.

The distinction between *ex ante* and *ex post* assessments demonstrate that different types of impacts come into focus at various moments in any impact assessment process, and that impacts can only be artificially bounded. The impacts that are discernible at the design and specification phase of a project are different than the impacts that become visible in other phases, particularly for ADSs that continuously patch, update, and scale. For that reason, *ex ante* assessments may be most useful as a form of transparency for technical or historical records. Since algorithmic systems also need to be assessed in terms of how interpretable or explainable their outputs are to human users, having a record of choices made in design (*ex ante*) are prerequisites for any forensic (*ex post*) investigation (Selbst and Barocas 2018). To fully understand impacts that produce harm to people requires careful consideration about when it becomes possible to anticipate, detect, and mitigate such harms.

The logistical question of when an impact is assessed during an AIA lays the groundwork for considering the larger ethical question of when is an impact. An impact is not just an articulation of a potential outcome with predefined attributes frozen in time—rather an outcome becomes an impact in practice, for someone, when it is connected to a form of assessment. An impact is a potential outcome that emerges in practices of assessment, connected to techniques of measurement and power relationships between the actors and the forum. As actionable objects that map the possible future(s) of ADSs, impacts forestalled are still impacts assessed. This way of approaching impacts, however, does not capture how impacts structure ambiguities of

[5] The prior knowledge necessary to anticipate, measure, and mitigate impacts is not without contention, as even baseline data about specific environmental quality measurements can be manipulated through the environmental impact assessment process. See Kinchy, Abby. 2020. "Contentious Baselining: The Politics of 'Pre-Drilling' Environmental Measures in Shale Gas Territory." Environment and Planning E: Nature and Space 3 (1): 76–94.

relations between the actors and the forum. The ethics of impact assessment is not only a matter of building standardized techniques of measuring impacts fairly and transparently, but also a matter of organizing relations. We explore this challenge in greater detail in our next observation.

Impact Assessments Structure Interactions in and between Institutions

Impact assessments bring different sets of organizations into formalized relationships with each other. These relationships have economic and political consequences. For instance, the structures and relationships that are established also set the conditions for different types of accountability. Assessment statutes create frameworks within which policymakers and technical actors are constrained and empowered when it comes to the design and implementation of a particular system (Solow-Niederman et al. 2019). Furthermore, some impact assessment regimes establish a public process by which different (often adversarial) actors, including the general public, are formally brought into dialogue. Other impact assessment regimes necessitate ongoing interaction between actors in ways that establish more collaborative rather than adversarial modes of interaction.

According to Serge Taylor's analysis of the Environmental Impact Assessment (EIA) process under the 1969 National Environmental Protection Act (NEPA), the EIA process places environmental advocacy organizations into an adversarial relationship with project developers through formal bureaucratic procedures within the Environmental Protection Agency (EPA) (Taylor 1984). The EIA process places developers, bureaucrats, environmental analysts, and advocacy organizations into a specific set of relations by requiring a proposed development plan be assessed by experts according to established guidelines before a project can move forward. The human rights impact assessment (HRIA) process, however, places institutions into vastly different relationships. Human rights experts are contracted by a corporate entity to produce an analysis of their business activities, and that analysis serves as a knowledge base from which that corporate actor may make voluntary choices to address potential human rights impacts within their control.

Different regimes of impact assessment, therefore, evoke specific forms of social and political power—between bureaucrats, developers, and public advocates, or between businesses and those whose human rights are impacted by business activities—that must be properly interrogated to scope a new impact assessment process. Nevertheless, over time an impact assessment regime can shift as new actors (agency departments, consulting companies, professional roles) respond to the demand for the work needed to complete impact assessments. New economies of compliance are created, and new entities can arise to take on duties that were intended to be performed by others, as with environmental consulting firms for the EIA process. As decision-making power shifts, so too does the locus of power, and nature of accountability. For example, much of the meaning and intent of compliance can shift towards powerful actors, when firms gradually take on the work of compliance for themselves and are only required to attest to their self-regulation to an oversight agency.

The exercise of institutional power leads to questions about who exerts power over whom, what checks on that power can be exercised, where resistance to power can be located, and how abuses of power can be ameliorated. Given the tendency for compliance practices to trend towards self-regulation, and in the absence of structural arrangements between powerful institutions that limit each other's power, the questions of the organizational values guiding impact assessment must be foregrounded. The onus of ethical behavior toward a broader public is placed squarely on private actors, who are often unaccountable to those who interact with their products. Oftentimes, this stands in stark contrast to the public interest, notions of a greater good, or even responsibility for harms that do not neatly fall under product liability standards. Therefore, accountability for organizations acting independently of other institutions remains attenuated.

In designing a new impact assessment process, particularly for AIAs, how relationships between organizations are structured is an important point of leverage that ought to be the subject of deliberation before formalizing them as a regulatory requirement. The existing technological development process typically involves documentation, and adding impact assessment-related specifications to existing documentation processes could be minimally disruptive, although this might differ between startups and more mature organizations. An important challenge for establishing accountability relationships in algorithmic development processes is the modular configuration of these technologies in which components produced by one company might be used by another and sold to a third without clear lines of demarcation around who is responsible for which components. Adding to this challenge, agile production methods exacerbate the problem of meaningfully stabilizing a system to conduct an AIA (Raji et al. 2020). Therefore, nailing down accountability relationships will remain a moving target. Understanding how relationships have been structured by other types of impact assessment and documentation processes will be crucial for deliberating over AIAs.

Public Participation in an Assessment Process is Not Synonymous with Accountability to the Public

Public participation is critical for democratic governance. In making rules for operation of federal agencies, it is a key mechanism for making the government more responsive and accountable to the public (Rowe and Frewer 2000). In environmental decision making, for example, public participation plays a strong role in education and resolving issues of conflict and mistrust (Beierle and Cayford 2010). The commenting process, further, can change an agency's course of action (Kochan 2017)

Additionally, the legitimacy of the impact assessment process depends on some degree of participation from a variety of stakeholders, including government agencies, private companies, consulting firms, and advocacy groups, as well as some definition of what constitutes "the public".[6] There have been similar calls for public participation in a wide range of AI policy documents "as a way to increase diversity, representation and equality in AI development and use" (Ulnicane et al. 2020).

However, not all forms of participation are equal. Different types of impact assessment mobilize different forms of representation and participation from respective constituencies through comment periods, focus groups, rapid assessments, or open meetings (Involve 2005, Fung 2015). How participation is defined or measured as "successful" are deeply contested issues (Rosener 1978, Ulnicane et al. 2020). The involvement of "stakeholders" is often a key component of public participation. However, scholarship in stakeholder theory finds that "stakeholders" are identified differently across institutions, with disparate definitions contingent on a social group's power to influence them, the legitimacy of a group's relationship with them, and the urgency and recognition of a group's claim (Mitchell et al. 1997). Efforts to bring stakeholders, however defined, together can engender new spaces of deliberation, collaboration, and empowerment (Young et al. 2019, Costanza-Chock 2020, Martin et al. 2020). However, they can also inadvertently flatten asymmetries in agency, power, voice, and vulnerability.

Moreover, despite the best of intentions, the relationship between public participation, transparency, and accountability, is far from straight-forward (Fox 2007). While critically important to a functioning and accountable democracy, public participation is not a panacea for the potential negative impacts of algorithmic systems, with recent scholarship calling attention to the risks of "participation-washing" in machine learning development (Sloane et al. 2020). Poorly designed commenting procedures can be easily gamed by actors seeking to discredit their validity (Grimaldo 2018). A lack of rigor and reflexivity in organizing for participation risks the possibility that they: (1) become a performance of caring for those who might be impacted, or (2) enroll vulnerable populations into harmful processes, or (3) make a community's vulnerability legible to bad actors.

The ethics of using public participation as a resource to encourage deliberation over possible consequences of algorithmic systems must consider the possibility that a preoccupation with procedure may render other important aspects of involving the public invisible. Efforts to encourage public participation without critical reflection on how they may also produce exploitative and extractive forms of community involvement can jeopardize public trust (Sloane et al. 2020). In terms of John Dewey's pragmatist philosophy, "the prime difficulty [of organizing public participation] is that of discovering the means by which a scattered, mobile and manifold public may so recognize itself as to define and express its interests" (Dewey 1927: 146). Publics need issues to gather around. For pragmatist philosophers, "the settlement of public issues depends on institutional outsiders adopting and articulating those issues, and bringing them to the attention of institutions that are equipped to deal with them." (Marres 2007: 775). Along these lines, ADSs have emerged as public issues through notable third-party audits that have driven attention to algorithmic harms and motivated increased adoption of internal auditing mechanisms. Notable examples include ProPublica's analysis of the Northpointe COMPAS recidivism prediction algorithm (led by Julia Angwin), the Gender Shades project's analysis of race and gender bias in facial recognition APIs offered by multiple companies (led by Joy Buolamwini), and Virginia Eubanks' account of algorithmic decision systems employed by social service agencies (Angwin et al. 2016, Buolamwini and Gebru 2018, Eubanks 2018). However, it is not self-evident that such third-party audits engender public involvement in governing ADSs writ large, rather they point to the need for more attention to a broad range of events in which algorithmic harms become widespread issues of public concern. Analyzing such events must begin with the understanding that while there are occasional events when issues engender a public and widespread mobilization of resources to address them, in most cases this does not happen. Identifying these occasions and describing how they create unique affordances for public involvement is a step towards achieving accountability of ADSs to the public. As we elaborate further in our next observation, these occasions can arise despite having impact assessment procedures in place.

Assessing Impacts Does Not Necessarily Mean Addressing Harms

An impact assessment itself does nothing to mitigate or directly address identified harms, although some assessment processes require mitigation of impacts to be explicitly documented. Rather, impact assessments provide information upon which other interventions or processes can build. Without identifying what impacts are, or what they are likely to be, it is impossible to mitigate harmful impacts, or govern a response to those impacts—and ultimately, to hold responsible parties accountable for those impacts. For most extant impact assessment processes, a great deal of attention has been paid to methodologies that can provide a knowledge base on which properly empowered actors can engage in rational decision-making. What constitutes a rationale for decision-making in the context of particular impact assessment regimes is an extension of how a particular form of impact assessment is imagined to facilitate further decision-making. Human rights impact assessments imagine corporate actors as willing to make changes to their business practices following an assessment and furthermore provide a mechanism

[6] See Jonathan Poisner, A Civic Republican Perspective on the National Environmental Policy Act's Process for Citizen Participation, 26 ENVTL. L. 53, 55 (1996) ("[C]itizen participation in the creation of NEPA-mandated [EISs] has, in all likelihood, spawned the largest amount of citizen participation in environmental decision making over the last two decades.").

for remedy for individuals who have been harmed. EIAs imagine impacts to environmental resources can be anticipated in advance of a development project so that less impactful design choices can be made or mitigation efforts can be mandated.

In order to satisfy the sometimes-competing goals of developers, government agencies, and advocates to undertake projects while limiting harmful impacts, there are necessarily trade-offs between known impacts and the overall benefits of an undertaking to society. Understanding how to assess the scope, scale, and depth of an impact will be necessary for understanding when a potential impact is acceptable, within the context of a given project. Understanding when a project must be altered or abandoned, and how to go about enforcing the needed changes, is crucial for any impact assessment process to fully realize accountable governance centering those who are most likely to be impacted by development projects.

Similar to other forms of impact assessment, any rigorous AIA will likely detect harms that go unremedied, but the overall process should be able to facilitate robust, engaged, and transparent decision-making around what the tradeoffs are between potential harms and likely benefits. The process through which harms are measured as impacts is an important component of the responsible development of ADSs but is ultimately in service of the pragmatic consequences of these systems on the world. It is in this calculation of tradeoffs that impact assessments can fall short of their promise to reduce harms and promote justice. Indeed, the discourse of cost-benefit analysis can justify diffuse benefits for many while excusing severe harms to a few. The impact assessment process itself can be a mechanism through which these machinations can be advanced by determining local harms are acceptable in the context of more widespread benefits—a calculation that has led to inadequate attention on environmental impacts to the minority-majority communities closest to unhealthy development projects (Cole 1992) and a failure to consider alternatives that would address historic environmental injustices (Bullard 1999).

Impact Assessments Ask Us How the World Might Be Otherwise

Impact assessments, by drawing attention to design choices and consequences, prompt a consideration of alternatives. They fundamentally rely on counterfactual reasoning: how might a resource be impacted if a project is undertaken, or a system implemented. Or, if considering an already deployed project or system: how did this project or system change the resource? By creating room for such alternatives in development cycles, impact assessments can shape bureaucratic or corporate decision-making, potentially leading to different and more thoughtful design choices.

Precisely because impact assessments have this counterfactual component, they create an opportunity to reorganize power when the constitutive components are configured to accomplish it. Impact assessments have the potential to provide a lever of influence for figures who may not otherwise hold power to shape policy, whether by providing opportunities to contest a proposal in advance or creating records useful in litigation after torts occur. Impact assessments might be a means for such communities to highlight otherwise overlooked or unforeseen sets of causes and effects ("What is Impact Assessment?" 2014). This is the approach that motivated the creation of the National Environmental Protection Act (NEPA). Drafters of NEPA chose not to pursue traditional methods of reform, like introducing additional legislation or engaging in drawn out political battles over agency leadership, and instead hoped that requiring agencies to make their environmental impacts transparent to the public would produce changes in how projects develop (Taylor 1984, Cheney-Lippold 2011, Hutchinson et al. 2020, Mohamed et al. 2020).

This feature of impact assessments is resonant with a central, if often overlooked, ethical consequence of machine learning. *The models at the core of an algorithmic system are functionally proposals for how the world ought to be, made concrete through deployment and integration of that system with existing sociotechnical arrangements.* They promise, for example, more efficient allocation of state resources, better rates of disease diagnosis, and optimized traffic flows. As machine learning model development moves from data lake to dataset, to sandbox, to production model, to deployed inside a product, each step becomes both a more economically valuable abstraction and a hardened set of preferences about how the world should be. Theorists of AI/ML systems have approached this world-building capacity of predictive models from a variety of standpoints (Cheney-Lippold 2011, Gebru et al. 2018, Mitchell et al. 2019, Eckersley 2019, Hutchinson et al. 2020, Mohamed et al. 2020, Raji et al. 2020), and many internal auditing mechanisms can be construed as methods for accounting how the engineering decisions behind a product may ultimately affect the world (Gebru et al. 2018, Mitchell et al. 2019, Hutchinson et al. 2020, Raji et al. 2020). Impact assessments propose to make visible how the world might be changed by a specific project—algorithmic or otherwise.

This counterfactual process is where the ethical stakes of our other observations come together. Who gets to enter this space, whose interests must be formally considered, which expertise and whose epistemology is treated as relevant to knowing this world, and when the counterfactual is considered to be adequately mapped are all sites of contestation and social power. Although AI systems have abstract mathematical constructs at their center, they create worlds that have actual consequences for people. AIAs, if constructed properly, can function as an obligation to comprehensively ask whether those worlds are desirable and for whom.

Conclusion

As policy makers and industry actors develop AI governance, it is crucial to remember that every governance structure will have benefits and drawbacks, and the devil is often in the proverbial

details. As we have outlined in these challenges, impact assessments encompass a wide range of approaches, methodologies, and opportunities. There is no universal path to follow. These challenges also point us towards the need for empirical research and social-science methodologies to better inform that which assessments are intended to assess, as well as how assessment practices intersect with other social processes in particular contexts, from economic development to the administration of justice to the cultural significance of demographic categories. Doing so would extend the original intention of an impact assessment regime to introduce grounded, empirical science to policy decision making (Taylor 1984).

AIAs in particular hold many challenges. On one hand, the algorithmic development process already presents several steps that could serve as ready-made handles for an impact assessment process to grab onto. Data collection design, data cleaning, model evaluation, and model deployment all represent moments when metadata relevant to the potential impacts of an ADS can be documented for the assessment process. These are also moments when interventions might be made to mitigate any potentially harmful impacts prior to deployment. On the other hand, there remains a great deal of ambiguity around how impacts are defined (and by whom), how they are assessed (and by whom), and how this establishes or fails to establish robust forms of accountability.

Acknowledgements

The authors wish to thank the many who reviewed and provided comments throughout the writing of this paper, including anonymous reviewers. The authors particularly wish to thank Patrick Davison, Michele Gilman, Mark Latonero, and Andrew Selbst for their intellectual generosity, the Raw Materials Seminar at Data & Society for the excellent and useful feedback, and the communications and editorial staff at Data & Society for their invaluable support. Portions of this work were made possible through the generous support of the NSF (awards #1704369 and #1633400), the Wenner Gren Foundation, and Luminate.

References

[1] Ada Lovelace Institute and DataKind UK, 2020. Examining the Blackbox: Tools for assessing algorithmic systems. https://www.adalovelaceinstitute.org/wp-content/uploads/2020/04/Ada-Lovelace-Institute-DataKind-UK-Examining-the-Black-Box-Report-2020.pdf

[2] Algorithmic Accountability Act. 2019. H.R. 2231, 116th Congress. (2019-2020).

[3] Allison-Hope, D., Darnton, H., and Lee, M. 2019. "Google's Human Rights by Design." Business for Social Responsibility blog: https://www.bsr.org/en/our-insights/blog-view/google-human-rights-impact-assessment-celebrity-recognition.

[4] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. Machine Bias. ProPublica.

[5] Bailey, P.D., Haq, G. and Gouldson, A., 2002. "Mind the gap! Comparing ex ante and ex post assessments of the costs of complying with environmental regulation." *European Environment*, *12*(5), pp.245-256.

[6] Beierle, T.C. and Cayford, J, 2010. Democracy in practice: Public participation in environmental decisions. Routledge.

[7] Bijker. W. 1995. Of bicycles, bakelites, and bulbs: toward a theory of sociotechnical change. MIT Press, Cambridge, MA.

[8] Bijker, W. Hughes, T., and Pinch, T. (Eds.). 1987. The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology. MIT Press, Cambridge, Mass.

[9] Bovens, M. 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

[10] Bullard, Robert D. 1999. "Anatomy of Environmental Racism and the Environmental Justice Movement." In Confronting Environmental Racism: Voices From the Grassroots, edited by Robert D. Bullard. South End Press.

[11] Buolamwini, J. and Gebru, T., 2018, January. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91).

[12] Cath, Corrine. 2018. "Governing artificial intelligence: ethical, legal and technical opportunities and challenges." Phil. Trans. R. Soc. A 376: 20180080.

[13] Cheney-Lippold, J. 2011. "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control." Theory, Culture & Society 28 (6): 164–81

[14] Cole, L.W. 1992. "Remedies for Environmental Racism: A View from the Field." *Michigan Law Review* 90, no. 7. https://doi.org/10.2307/1289740.

[15] Costanza-Chock, S. 2020. Design Justice: Community-led Practices to Build the Worlds We Need. MIT Press: Cambridge, MA.

[16] Datta, A., Anupam D., Makagon, J., Mulligan, D., and Tschantz, M.C. 2018. "Discrimination in Online Advertising: A Multidisciplinary Inquiry." In Conference on Fairness, Accountability and Transparency, 20–34. PMLR. http://proceedings.mlr.press/v81/datta18a.html

[17] DeLong, J.V. 1979. "Informal Rulemaking and the Integration of Law and Policy." Virginia Law Review 65, no. 2: 257-356.

[18] Dewey, J. 1927. *The Public and Its Problems*. New York: H. Holt and Company.

[19] Eckersley, P. 2019. "Impossibility and Uncertainty Theorems in AI Value Alignment (or Why Your AGI Should Not Have a Utility Function)." ArXiv:1901.00064 [Cs], March. http://arxiv.org/abs/1901.00064.

[20] Elish, M.C. 2019. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." Engaging Science, Technology, and Society 5 (0): 40–60. https://doi.org/10.17351/ests2019.260

[21] Elish, M.C. and Watkins, E.A. 2020. "Repairing Innovation: A Study of Integrating AI in Clinical Care." Data & Society Research Institute.

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

[23] Fox, Jonathan. 2007. "The uncertain relationship between transparency and accountability." *Development in Practice*, 17:4, 683-671.

[24] Fung, Archon. 2015. "Putting the Public Back into Governance: The Challenges of Citizen Participation and Its Future." Public Administration Review 75(4), pp.513-522

[25] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H. and Crawford, K., 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

[26] Government of Canada. 2019. Directive on Automated Decision-Making. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

[27] Grimaldo, J. 2018. "New York Attorney General's Probe Into Fake FCC Comments Deepens." Wall Street Journal.

[28] Hanna, A., and Park, T.M. 2020. "Against Scale: Provocations and Resistances to Scale Thinking." ArXiv:2010.08850 [Cs], October. http://arxiv.org/abs/2010.08850.

[29] Hoffmann, A.L. 2020. "Terms of Inclusion: Data, Discourse, Violence." New Media & Society, September, 1461444820958725.

[30] Holland, S., Hosny, A., Newman, S. Joseph, J., Chmielinski, K. 2018. The dataset nutrition label: A framework to drive higher quality data standards. arXiv:1805.03677

[31] Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. 2020. "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure." ArXiv:2010.13561 [Cs], October. http://arxiv.org/abs/2010.13561.

[32] Involve. 2005. People & Participation: How to put citizens at the heart of decision-making. Beacon Press: London. Available online: https://www.involve.org.uk/sites/default/files/uploads/People-and-Participation.pdf

[33] Jasanoff, S. 2004. States of knowledge: the co-production of science and the social order. Routledge.

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

[35] Kaminski, M. and Malgieri, G. 2019. "Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations." U of Colorado Law Legal Studies Research Paper No. 19-28. Available at SSRN: https://doi.org/10.2139/ssrn.3456224

[36] Karlin, M. and Corriveau, N. 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.

[37] Keddell, E. 2019. "Algorithmic Justice in Child Protection: Statistical Fairness, Social Justice and the Implications for Practice." Social Sciences 8 (10): 281. https://doi.org/10.3390/socsci8100281.

[38] Keyes, O. 2018. "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." Proceedings of the ACM on Human-Computer Interaction 2 (CSCW): 88:1-88:22. https://doi.org/10.1145/3274357.

[39] Kochan, D.J., 2017. The commenting power: Agency accountability through public participation. *Okla. L. Rev.*, *70*, p.601.

[40] Koene, Ansgar, Chris Clifton, Yohko Hatada, Helena Webb, and Rashida Richardson. 2019. "A governance framework for algorithmic accountability and transparency." Brussels: European Parliamentary Research Service.

[41] Latour, B. 2005. Reassembling the social: an introduction to actor-network theory. Oxford University Press, Oxford

[42] Lynch, M. 2001. The epistemology of epistopics: Science and technology studies as an emergent (non)discipline. *American Sociological Association, Science, Knowledge & Technology Section (ASA-SKAT) Newsletter* (Fall): 2–3. https://asaskat.files.wordpress.com/2015/11/skat-331.pdf

[43] Lynch, M. 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;

[44] Marres, N. 2007. "The Issues Deserve More Credit: Pragmatist Contributions to the Study of Public Involvement in Controversy." *Social Studies of Science* 37 (5): 759–80. https://doi.org/10.1177/0306312706077367.

[45] Martin Jr., D., Prabhakaran, V., Kuhlberg, J., Smart, A., & Issac, W. S. 2020. Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics. Fair & Responsible AI Workshop. CHI 2020.

[46] Mauer, M. 2007. "Racial Impact Statements as a Means of Reducing Unwarranted Sentencing Disparities." *Ohio State Journal of Criminal Law* 5: 28.

[47] Metcalf, J., Moss, E., Watkins, E.A., Singh, R., and Elish, M.C. 2021. "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts." In *Proceedings of the ACM Conference on Fairness, Accountability and Transparency* (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442188.3445935

[48] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T., 2019, January. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

[49] Mitchell, R.K., Agle, B.R. and Wood, D.J., 1997. "Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts." *Academy of management review*, *22*(4), pp.853-886.

[50] Mohamed, S., Png, M., and Isaac, W. 2020. "Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence." *Philosophy & Technology* 33 (4): 659–84. https://doi.org/10.1007/s13347-020-00405-8

[51] Moss, E., Watkins, E.A., Singh, R., Elish, M.C., and Metcalf, J. (forthcoming). "Assembling Accountability Through Algorithmic Impact Assessment." Data & Society Research Institute. http://datasociety.net/library/assembling-accountability/

[52] Mourey, J. and Waldman, A. 2020. "Past the Privacy Paradox: The Importance of Privacy Changes as a Function of Control and Complexity". *Journal of the Association for Consumer Research* 5:2, 162-180.

[53] Mulligan, D. K. and Kenneth A. B. 2019. "Procurement as Policy: Administrative Process for Machine Learning". Berkeley Technology Law Journal, Vol. 34, 2019. Available at SSRN: https://ssrn.com/abstract=3464203.

[54] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P., 2020, January. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44).

[55] Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. 2018. "Algorithmic impact assessments: A practical framework for public agency accountability." AI Now Institute. 1-22. https://ainowinstitute.org/aiareport2018.pdf.

[56] Rona-Tas, A. 2017. The Off-Label Use of Consumer Credit Ratings. Historical Social Research, 42(1 (159)), 52-76.

[57] Rosener, J. B. 1978. "Citizen participation: Can we measure its effectiveness?" *Public Administration Review*, September/October, 457-63.

[58] Rowe, G. and Frewer, L.J., 2000. Public participation methods: A framework for evaluation. *Science, technology, & human values*, 25(1), pp.3-29.

[59] Selbst, A. 2017. "Disparate Impact in Big Data Policing," 52 Georgia L. Rev., https://ssrn.com/ abstract=2819182

[60] Selbst, A. and Barocas, S., 2018. "The Intuitive Appeal of Explainable Machines." *87 Fordham Law Review 1085* (2018).

[61] Shapiro, D.L. 1965 "The Choice of Rulemaking or Adjudication in the Development of Administrative Policy." Harvard Law Review 78, no. 5 (1965): 921-72.

[62] Sloane, M., Moss, E., Awomolo, O. and Forlano, L., 2020. Participation is not a design fix for machine learning. Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria.

[63] Solow-Niederman, A., Choi, Y., and Van den Broeck, G. 2019. "Institutional Life of Algorithmic Risk Assessments." Berkeley Technology Law Journal 34 (705): 05–744.

[64] Taylor, S., 1984. Making bureaucracies think: The environmental impact statement strategy of administrative reform. *Making bureaucracies think: the environmental impact statement strategy of administrative reform.*CA: Stanford University Press

[65] Tufekci, Z., 2015. Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colo. Tech. LJ*, *13*, p.203.

[66] Ulnicane, I, Knight, W., Leach, T., Stahl, B.C., and Wanjiku, W. 2020. "Framing Governance for a Contested Emerging Technology: Insights from AI Policy." *Policy and Society*, December, 1–20. https://doi.org/10.1080/14494035.2020.1855800.

[67] Veale, M. and Brass, I. 2019 "Administration by Algorithm? Public Management Meets Public Sector Machine Learning." In: Algorithmic Regulation (Karen Yeung and Martin Lodge eds., Oxford University Press, 2019). Available at SSRN: https://ssrn.com/abstract=3375391.

[68] Wadhwa, K. and Rodrigues, R., 2013. Evaluating privacy impact assessments. *Innovation: The European Journal of Social Science Research*, *26*(1-2), pp.161-180.

[69] Waldman, A.E. 2019. "Privacy Law's False Promise". *Washington University Law Review*, Vol. 97, No. 2, 2020. Available at SSRN: https://ssrn.com/abstract=3499913.

[70] Waldman, A.E. 2020. "Cognitive Biases, Dark Patterns, and the 'Privacy Paradox.'" *Current Opinion in Psychology* 31: 105–9.

[71] Warofka, A. 2018. "An Independent Assessment of the Human Rights Impact of Facebook in Myanmar." Facebook press release, Nov 5: https://about.fb.com/news/2018/11/myanmar-hria/

[72] West, W. 2005. "Administrative Rulemaking: An Old and Emerging Literature." Public Administration Review 65, no. 6: 655-68.

[73] "What is Impact Assessment?" 2014. Based on OECD Directorate for Science, Technology and Innovation, "Assessing the Impact of State Interventions in Research – Techniques, Issues and Solutions", unpublished manuscript.

[74] Wieringa, M. 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, 1–18. Barcelona Spain: ACM. https://doi.org/10.1145/3351095.3372833.

[75] Wright, D., 2011. Should privacy impact assessments be mandatory?.*Communications of the ACM*, *54*(8), pp.121-131.

[76] Wright, D. and Friedewald, M., 2013. Integrating privacy and ethical impact assessments. Science and Public Policy, 40(6), pp.755-766.

[77] Young, Meg, Lassana Magassa, and Batya Friedman. 2019. "Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents." *Ethics and Information Technology 21:* 89-103.

[78] Ziewitz, M. 2019. Rethinking gaming: The ethical work of optimization in web search engines. *Social Studies of Science*. 49(5):707-731. doi:10.1177/0306312719865607