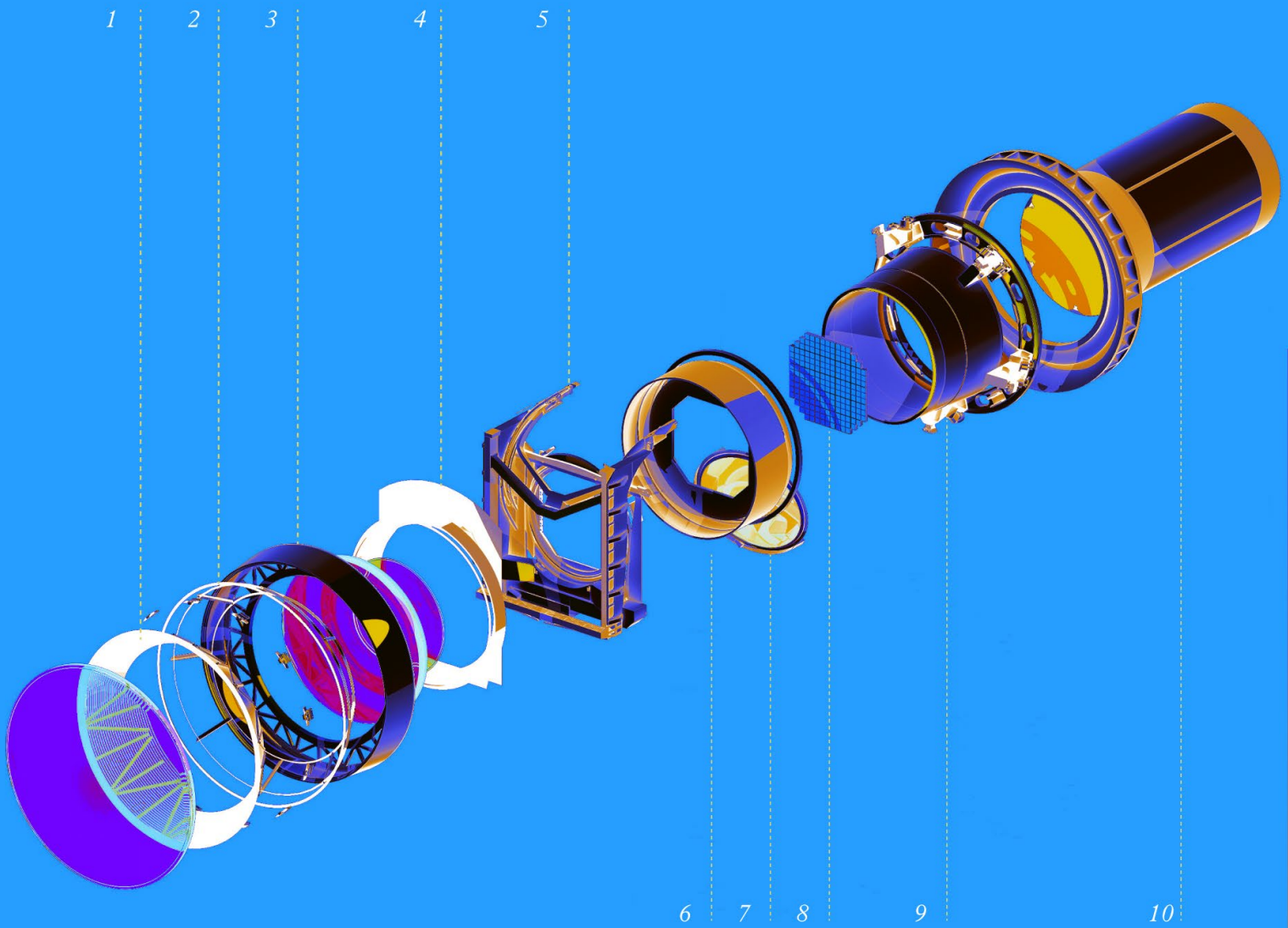
Assembling Accountability

Algorithmic Impact Assessment for the Public Interest

Emanuel Moss
Elizabeth Anne Watkins
Ranjit Singh
Madeleine Clare Elish
Jacob Metcalf

DATA&
SOCIETY



SUMMARY

This report maps the challenges of constructing algorithmic impact assessments (AIAs) by analyzing impact assessments in other domains—from the environment to human rights to privacy. Impact assessment is a promising model of algorithmic governance because it bundles an account of potential and actual harms of a system with a means for identifying who is responsible for their remedy. Success in governing with AIAs requires thoughtful engagement with the ongoing exercise of social and political power, rather than defaulting to self-assessment and narrow technical metrics. Without such engagement, AIAs run the risk of not adequately facilitating the measurement of, and contestation over, harms experienced by people, communities, and society.

We use existing impact assessment processes to showcase how "impacts" are evaluative constructs that create the conditions for diverse institutions—private companies, government agencies, and advocacy organizations—to act in response to the design and deployment of systems. We showcase the necessity of attending to how impacts are constructed as meaningful measurements, and analyze occasions when the impacts measured do not capture the actual on-the-ground harms. Most concretely, we identify 10 constitutive components that are common to all existing types of impact assessment practices:

1) Sources of Legitimacy,
2) Actors and Forum,
3) Catalyzing Event,
4) Time Frame,
5) Public Access,
6) Public Consultation,
7) Method,
8) Assessors,
9) Impact,
10) Harms and Redress.

By describing each component, we build a framework for evaluating existing, proposed, and future AIA processes. This framework orients stakeholders to parse through and identify components that need to be added or reformed to achieve robust algorithmic accountability. It further underpins the overlapping and interlocking relationships between differently positioned stakeholders—regulators, advocates, public-interest technologists, technology companies,

and critical scholars—in identifying, assessing, and acting upon algorithmic impacts. As these stakeholders work together to design an assessment process, we offer guidance through the potential failure modes of each component by exploring the conditions that produce a widening gap between impacts-as-measured and harms-on-the-ground.

This report does not propose a specific arrangement of these constitutive components for AIAs. In the co-construction of impacts and accountability, what impacts should be measured only becomes visible with the emergence of who is implicated in how accountability relationships are established. Therefore, we argue that any AIA process only achieves real accountability when it:

a) keeps algorithmic "impacts" as close as possible to actual algorithmic harms;
b) invites a broad and diverse range of participants into a consensus-based process for arranging its constitutive components; and
c) addresses the failure modes associated with each component.

These features also imply that there will never be one single AIA process that works for every application or every domain. Instead, every successful AIA process will need to adequately address the following questions:

a) Who should be considered as stakeholders for the purposes of an AIA?
b) What should the relationship between stakeholders be?
c) Who is empowered through an AIA and who is not? Relatedly, how do disparate forms of expertise get represented in an AIA process?

Governing algorithmic systems through AIAs will require answering these questions in ways that reconfigure the current structural organization of power and resources in the development, procurement, and operation of such systems. This will require a far better understanding of the technical, social, political, and ethical challenges of assessing the value of algorithmic systems for people who live with them and contend with various types of algorithmic risks and harms.

CONTENTS

INTRODUCTION

The last several years have been a watershed for algorithmic accountability. Algorithmic systems have been used for years, in some cases decades, in all manner of important social arenas: disseminating news, administering social services, determining loan eligibility, assigning prices for on-demand services, informing parole and sentencing decisions, and verifying identities based on biometrics among many others. In recent years, these algorithmic systems have been subjected to increased scrutiny in the name of accountability through adversarial quantitative studies, investigative journalism, and critical qualitative accounts. These efforts have revealed much about the lived experience of being governed by algorithmic systems. Despite many promises that algorithmic systems can remove the old bigotries of biased human judgement,[1] there is now ample evidence that algorithmic systems exert power precisely along those familiar vectors, often cementing historical human failures into predictive analytics. Indeed, these systems have disrupted democratic electoral politics,[2] fueled violent genocide,[3] made vulnerable families even more vulnerable,[4] and perpetuated racial- and gender-based discrimination.[5]

Algorithmic justice advocates, scholars, tech companies, and policymakers alike have proposed algorithmic impact assessments (AIAs)—borrowing from the language of impact assessments from other domains—as a potential process for addressing algorithmic harms that moves beyond narrowly constructed metrics towards real justice.[6] Building an impact assessment process for algorithmic systems raises several challenges. For starters, assessing impacts requires assembling a multiplicity of viewpoints and forms of expertise. It involves deciding whether sufficient, reliable, and adequate amounts of evidence have been collected about systems' consequences on the world, but also about their formal structures—technical specifications, operating parameters, subcomponents, and ownership.[7] Finally, even when AIAs (in whatever form they may take) are conducted, their effectiveness in addressing on-the-ground harms remains uncertain.

1 Anne Milgram, Alexander M. Holsinger, Marie Vannostrand, and Matthew W. Alsdorf, "Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making," *Federal Sentencing Reporter* 27, no. 4 (2015): 216–21, https://doi.org/10.1525/fsr.2015.27.4.216. *cf.* Angèle Christin, "Algorithms in Practice: Comparing Web Journalism and Criminal Justice," *Big Data & Society* 4, no. 2 (2017): 205395171771885, https://doi.org/10.1177/2053951717718855.

2 Carole Cadwalladr, and Emma Graham-Harrison, "The Cambridge Analytica Files," *The Guardian*, https://www.theguardian.com/news/series/cambridge-analytica-files.

3 Alexandra Stevenson, "Facebook Admits It Was Used to Incite Violence in Myanmar," *The New York Times*, November 6, 2018, https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

4 Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018), https://www.amazon.com/Automating-Inequality-High-Tech-Profile-Police/dp/1250074312.

5 Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of Machine Learning Research*, Vol. 81 (2018), http://proceedings.mlr.press/v81/buolamwini18a.html.

6 Andrew D. Selbst, "Disparate Impact in Big Data Policing," *SSRN Electronic Journal*, 2017, https://doi.org/10.2139/ssrn.2819182; Anna Lauren Hoffmann, "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse," *Information, Communication & Society* 22, no. 7(2019): 900–915, https://doi.org/10.1080/1369118X.2019.1573912.

7 Helen Nissenbaum, "Accountability in a Computerized Society," *Science and Engineering Ethics* 2, no. 1 (1996): 25–42, https://doi.org/10.1007/BF02639315.

Critics of regulation, and regulators themselves, have often argued that the complexity of algorithmic systems makes it impossible for lawmakers to understand them, let alone craft meaningful regulations for them.[8] Impact assessments, however, offer a means to describe, measure, and assign responsibility for impacts without the need to encode explicit, scientific understandings in law.[9] We contend that the widespread interest in AIAs comes from how they integrate measurement and responsibility—an impact assessment bundles together an account of *what this system does* and *who should remedy its problems*. Given the diversity of stakeholders involved, impact assessments mean many different things to different actors—they may be about compliance, justice, performance, obfuscation through bureaucracy, creation of administrative leverage and influence, documentation, and much more. Proponents of AIAs hope to create a point of leverage for people and communities to demand transparency and exert influence over algorithmic systems and how they affect our lives. In this report we show that the choices made about an impact assessment process determine how, and whether, these goals are achieved.

Impact assessment regimes principally address three questions: what a system does; who can do something about what that system does; and who ought to make decisions about what the system is permitted to do. Attending to how AIA processes are assembled is imperative because they may be the means through which a broad cross-section of society can exert influence over how algorithmic systems affect everyday life. Currently, the contours of algorithmic accountability are underspecified. A robust role for individuals, communities, and regulatory agencies outside of private companies is *not* guaranteed. There are strong economic incentives to keep accountability practices fully internal to private corporations. In tracing how IA processes in other domains have evolved over time, we have found that the degree and form of accountability emerging from the construction of an impact assessment regime varies widely and is a result of decisions made during their development. In this report, we illustrate the decision points that will be critical in the development of AIAs, with a particular focus on protecting and empowering individuals and communities who are systemically vulnerable to algorithmic harms.

One of the central challenges to designing AIAs is what we call the *specification dilemma*: Algorithmic systems can cause harm when they fail to work as specified—i.e., in error—but may just as well cause real harms when working *exactly* as specified. A good example for this dilemma is facial recognition technologies. Harms caused by inaccuracy and/or disparate accuracy rates of such technologies are well documented. Disparate accuracy across demographic groups is a form of error, and produces harms such as wrongful arrest,[10] inability to enter

8 Mike Snider, "Congress and Technology: Do Lawmakers Understand Google and Facebook Enough to Regulate Them?" *USA TODAY*, August 2, 2020, https://www.usatoday.com/story/tech/2020/08/02/google-facebook-and-amazon-too-technical-congress-regulate/5547091002/.

9 Serge Taylor, *Making Bureaucracies Think: The Environmental Impact Statement Strategy of Administrative Reform* (Stanford, CA: Stanford University Press, 1984).

10 Kashmir Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match," *The New York Times*, December 29, 2020, https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

one's own apartment building,[11] and exclusion from platforms on which one earns income.[12] In particular, false arrests facilitated by facial recognition have been publicly documented several times in the past year.[13] On such occasions, the harm is not merely the error of an inaccurate match, but an ever-widening circle of consequences to the target and their family: wrongful arrest, time lost to interrogation, incarceration and arraignment, and serious reputational harm.

Harms, however, can also arise when such technologies are working as designed.[14] Facial recognition, for example, can produce harms by chilling rights such as freedom of assembly, free association, and protections against unreasonable searches.[15] Furthermore, facial recognition technologies are often deployed to target minority communities that have already been subjected to long histories of surveillance.[16] The expansive range of potential applications for facial recognition presents a similar range of its potential harms, some of which fit neatly into already existing taxonomies of algorithmic harm,[17] but many more of which are tied to their contexts of design and use.

Such harms are simply not visible to the narrow algorithmic performance metrics derived from technical audits. Another process is needed to document *algorithmic harms*, allowing: (a) developers to redesign their products to mitigate known harms; (b) vendors to purchase products that are less harmful; and (c) regulatory agencies to meaningfully evaluate the tradeoff between benefits and harms of appropriating such products. Most importantly, the public—particularly vulnerable individuals and communities—can be made aware of the possible consequences of such systems. Still, anticipating algorithmic harms can be an unwieldy task for any of these stakeholders—developers, vendors, and regulatory authorities—individually. Understanding algorithmic harms requires a broader community of experts: community advocates, labor organizers, critical scholars, public interest technologists, policy

11 Tranae' Moran, "Atlantic Plaza Towers Tenants Won a Halt to Facial Recognition in Their Building: Now They're Calling on a Moratorium on All Residential Use," *AI Now Institute* (blog), January 9, 2020, https://medium.com/@AINowInstitute/atlantic-plaza-towers-tenants-won-a -halt-to-facial-recognition-in-their-building-now-theyre-274289a6d8eb.

12 John Paul Brammer, "Trans Drivers Are Being Locked Out of Their Uber Accounts," *Them*, August 10, 2018, https://www.them.us/story/trans-drivers-locked-out-of-uber.

13 Bobby Allyn, "'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man," *NPR*, June 24, 2020, https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong- how-facial-recognition-led-to-a-false-arrest-in-michigan.

14 Commercial facial recognition applications like Clearview AI, for example, have been called "a nightmare for stalking victims" because they let abusers easily identify potential victims in public, and heighten the fear among potential victims merely by existing. Absent any user controls to prevent stalking, such harms are seemingly baked into the business model. See, for example, Maya Shwayder, "Clearview AI Facial-Recognition App Is a Nightmare For Stalking Victims," *Digital Trends*, January 22, 2020, https://www.digitaltrends.com/news/clearview-ai-facial-recognition-domestic-violence-stalking/; and Rachel Charlene Lewis, "Making Facial Recognition Easier Might Make Stalking Easier Too," *Bitch Media*, January 31, 2020, https://www.bitchmedia.org/article/very-online/clearview-ai-facial-recognition-stalking-sexism.

15 Kristine Hamann and Rachel Smith, "Facial Recognition Technology: Where Will It Take Us?" *Criminal Justice Magazine*, 2019, https://www.americanbar.org/groups/criminal_justice/publications/criminal-justice-magazine/2019/spring/facial-recognition-technology/.

16 Simone Browne, *Dark Matters: On the Surveillance of Blackness* (Durham, NC: Duke University Press, 2015).

17 Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach, "The problem with bias: from allocative to representational harms in machine learning," *Special Interest Group for Computing, Information and Society (SIGCIS) 2017.*

makers, and the third-party auditors who have been slowly developing the tools for anticipating algorithmic harms.

This report provides a framework for how such a diversity of expertise can be brought together. By analyzing existing impact assessments in domains ranging from the environment to human rights to privacy, this report maps the challenges facing AIAs.

Most concretely, we identify **10 constitutive components** that are common to all existing types of impact assessment practices (see table on page 50). Additionally, we have interspersed vignettes of impact assessments from other domains throughout the text to illustrate various ways of arranging these components. Although AIAs have been proposed and adopted in several jurisdictions, these examples have been constructed very differently, and none of them have adequately addressed all the **10 constitutive components**.

This report does not ultimately propose a specific arrangement of **constitutive components** for AIAs. We made this choice because impact assessment regimes are evolving, power-laden, and highly contested—the capacity of an impact assessment regime to address harms depends in part on the organic, community-directed development of its components. Indeed, in the co-construction of impacts and accountability, *what* impacts should be measured only becomes visible with the emergence of *who* is implicated in *how* accountability relationships are established.

We contend that the timeliest need in algorithmic governance is establishing the methods through which robust AIA regimes are organized. If AIAs are to prove an effective model for governing algorithmic systems, and, most importantly, protect individuals and communities from algorithmic harms, then they must:

a) keep algorithmic "impacts" as close as possible to actual algorithmic harms;
b) invite a diverse range of participants into the process of arranging its constitutive components; and
c) overcome the failure modes of each component.

WHAT IS AN IMPACT?

No existing impact assessment process provides a definition of "impact" that can be simply operationalized by AIAs. **Impacts** are *evaluative constructs* that enable institutions to coordinate action in order to identify, minimize, and mitigate harms. By evaluative constructs, we mean that impacts are not prescribed by a system; instead, they must be defined, and defined in a manner than can be measured. Impacts are *not* identical to harms: an impact might be disparate error rates for men and women within a hiring algorithm; the harm would be unfair exclusion from the job. Therefore, effective impact assessment requires identifying harms before determining how to measure impacts, a process which will differ across sectors of algorithmic systems (e.g., biometrics, employment, financial, et cetera).[18]

18 Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish, "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 735–746. FAccT '21*, (New York, NY, USA: Association for Computing Machinery, 2021), https://doi.org/10.1145/3442188.3445935.

Conceptually, "impact" implies a causal relationship: an action, decision, or system causes a change that affects a person, community, resource, or other system. Often, this is expressed as a counterfactual where the impact is the difference between two (or more) possible outcomes—a significant aspect of the craft of impact assessment is measuring "how might the world be otherwise if the decisions were made differently?"[19] However, it is difficult to precisely identify causality with impacts. This is especially true for algorithmic systems whose effects are widely distributed, uneven, and often opaque. This inevitably raises a two-part question: what effects (harms) can be identified as impacts resulting from or linked to a particular cause, and how can that cause be properly attributed to a system operated by an organization?

Raising these questions together points to an important feature of "impacts": *Harms are only made knowable as "impacts" within an accountability regime which makes it possible to assign responsibility for the effects of a decision, action, or system.* Without accountability relationships that delimit responsibility and causality, there are no "impacts" to measure; without impacts as a common object to act upon, there are no accountability relationships. Impacts, thus, are a type of *boundary object*, which, in the parlance of sociology of science, indicates a constructed or shared object that enables inter- and intra-institutional collaboration precisely because it can be described from multiple perspectives.[20] Boundary objects render a diversity of perspectives into a source of productive friction and collaboration, rather than a source of breakdown.[21]

For example, consider environmental impact assessments. First mandated in the US by the National Environmental Protection Act (NEPA) (1970), environmental impact assessments have evolved through litigation, legislation, and scholarship to include a very broad set of "impacts" to diverse environmental resources. Included in an environmental impact statement for a single project may be chemical pollution, sediment in waterways, damage to cultural or archaeological artifacts, changes to traffic patterns, human population health consequences, loss of habitat for flora and fauna, and a consideration of how (in)equitably environmental harms have been distributed across local communities in the past.[22] Such a diversity of measurements would not typically be grouped together; there are too many distinct methodologies and types of expertise involved. However, the accountability regimes that have evolved from NEPA create and maintain a conceptual and organizational framework that enables institutions to come together around a common object called an "environmental impact."

19 Matthew Cashmore, Richard Gwilliam, Richard Morgan, Dick Cobb, and Alan Bond, "The Interminable Issue of Effectiveness: Substantive Purposes, Outcomes and Research Challenges in the Advancement of Environmental Impact Assessment Theory," *Impact Assessment and Project Appraisal* 22, no. 4 (2004): 295–310, https://doi.org/10.3152/147154604781765860.

20 Susan Leigh Star, and James R. Griesemer, "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39," *Social Studies of Science* 19, no. 3 (1989): 387–420, https://doi.org/10.1177/030631289019003001; and Susan Leigh Star, "This Is Not a Boundary Object: Reflections on the Origin of a Concept," *Science, Technology, & Human Values* 35, no. 5 (2010): 601–17, https://doi.org/10.1177/0162243910377624.

21 Unlike other prototypical boundary objects from the science studies literature, impacts are centered on accountability, rather than practices of building shared scientific ontologies.

22 Judith Petts, *Handbook of Environmental Impact Assessment Volume 2: Impact and Limitations*. Vol. 2. 2 vols. (Oxford: Blackwell Science, 1999); Peter Morris, and Riki Therivel, *Methods of Environmental Impact Assessment* (London; New York: Spon Press, 2001), http://site.ebrarxy.com/id/5001176.

Impacts and accountability are *co-constructed*; that is, impacts do not precede the identification of responsible parties. What might be an impact in one assessment emerges from which parties are being held responsible, or from a specific methodology adopted through a consensus-building process among stakeholders. The need to address this co-construction of accountability and impacts has been neglected thus far in AIA proposals. As we show in existing impact assessment regimes, the process of identifying, measuring, formalizing, and accounting for "impacts" is a power-laden process that does not have a neutral endpoint. Precisely because these systems are complex and multi-causal, defining what counts as an impact is contested, shaped by social, economic, and political power. For all types of impact assessments, the list of impacts considered assessable will necessarily be incomplete, and assessments will remain partial. The question at hand for AIAs, as they are still at an early stage, is what are the standards for deciding when an AIA is complete enough?

WHAT IS ACCOUNTABILITY?

If impacts and accountability are co-constructed, then carefully defining accountability is a crucial part of designing the impact assessment process. A widely used definition of accountability in the algorithmic accountability literature is taken from a 2007 article by sociologist Mark Bovens, who argues that **accountability is "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences."**[23] Building on Boven's general articulation of accountability, Maranka Wierenga describes algorithmic accountability as "a networked account for a socio-technical algorithmic system, following the various stages of the system's lifecycle," in which "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."[24]

Following from this definition, we argue that **voluntary commitments to auditing and transparency do not constitute accountability.** Such commitments are not ineffectual—they have important effects, but they do not meet the standard of accountability to an external forum. They remain internal to the set of designers, engineers, software companies, vendors, and operators who already make decisions about algorithmic systems; *there is no distinction between the "actor" and the "forum."* This has important implications for the emerging field of algorithmic accountability, which has largely focused on technical metrics and internal platform governance mechanisms. While the technical auditing and metrics that have come out of the algorithmic fairness, accountability, transparency scholarship, and research departments of technology companies would inevitably constitute the bulk of an assessment process, *without an external forum such methods cannot achieve genuine accountability.* This in turn points to an underexplored dynamic in algorithmic governance that is the heart of this report: how should the measurement of algorithmic impacts be coordinated through institutional

23 Mark Bovens, "Analysing and Assessing Accountability: A Conceptual Framework," *European Law Journal* 13, no. 4 (2007): 447–68, https://doi.org/10.1111/j.1468-0386.2007.00378.x.

24 Maranke Wieringa, "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* (2020), 1–18, https://doi.org/10.1145/3351095.3372833.

practices and sociopolitical contestation to reduce algorithmic harms? In other domains, these forces and practices have been co-constructed in diverse ways that hold valuable lessons for the development of any incipient algorithmic impact assessment process.

WHAT IS IMPACT ASSESSMENT?

Impact assessment is a process for simultaneously documenting an undertaking, evaluating the impacts it might cause, and assigning responsibility for those impacts. Impacts are typically measured against alternative scenarios, including scenarios in which no development occurs. These processes vary across domains; while they share many characteristics, each impact assessment regime has its own historically situated approach to constituting accountability. Throughout this report, we have included short narrative examples for the following five impact assessment practices from other domains[25] as sidebars:

1. **Fiscal Impact Assessments (FIA)** are analyses meant to bridge city planning with local economics, by estimating the fiscal impacts such as potential costs and revenues that result from developments. Changes resulting from new developments as captured in the resulting report can include local employment, population, school enrollment, taxation, and other aspects of a government's budget.[26] *See page 12.*

2. **Environmental Impact Assessments (EIA)** are investigations that make legible to permitting agencies the evolving scientific consensus around the environmental consequences of development projects. In the United States, EIAs are conducted for proposed building projects receiving federal funds or crossing state lines. The resulting report might include findings about chemical pollution, damage to cultural or archaeological sites, changes to traffic patterns, human population health consequences, loss of habitat for flora and fauna, and/or a consideration of how (in)equitably environmental harms have been distributed across local communities in the past.[27] *See page 19.*

3. **Human Rights Impact Assessments (HRIA)** are investigations commissioned by companies or agencies to better understand the impact their operations—such as supply chain management, change in policy, or resource management—have on human rights, as defined by the Universal Declaration on Human Rights. Usually conducted by third-party firms and resulting in a report, these assessments ideally help identify and address the adverse effects

25 There are certainly many other types of impact assessment processes—social impact assessment, biodiversity impact assessment, racial equity impact assessment, health impact assessment—however, we chose these five as initial resources to build our framework of constitutive components because of similarity with some common themes of algorithmic harms and extant use by institutions that would also be involved in AIAs.

26 Zenia Kotval, and John Mullin, "Fiscal Impact Analysis: Methods, Cases, and Intellectual Debate," Lincoln Institute of Land Policy Working Paper, Lincoln Institute of Land Policy, 2006,https://www.lincolninst.edu/sites/default/files/pubfiles/kotval-wp06zk2.pdf.

27 Petts, *Handbook of Environmental Impact Assessment Volume 2*; Morris and Therivel, *Methods of Environmental Impact Assessment*.

of company or agency actions from the viewpoint of the rightsholder.[28] *See page 27.*

4. **Data Protection Impact Assessments (DPIA)**, required by the General Data Protection Regulation (GDPR) of private companies collecting personal data, include cataloguing and addressing system characteristics and the risks to people's rights and freedoms presented by the collection and processing of personal data. DPIAs are a process for both 1) building and 2) demonstrating compliance with GDPR requirements.[29] *See page 31.*

5. **Privacy Impact Assessments (PIA)** are a cataloguing activity conducted internally by federal agencies and, increasingly, companies in the private sector, when they launch or change a process which manages Personally Identifiable Information (PII). During a PIA, assessors catalogue methods for collecting, handling, and protecting PII they manage on citizens for agency purposes, and ensure that these practices conform to applicable legal, regulatory, and policy mandates.[30] The resulting report, as legislatively mandated, must be made publicly accessible. *See page 35.*

28 Mark Latonero, "Governing Artificial Intelligence: Upholding Human Rights & Dignity," Data & Society Research Institute, 2018, https://datasociety.net/library/governing-artificial-intelligence/; Nora Götzmann, Tulika Bansal, Elin Wrzoncki, Cathrine Poulsen-Hansen, Jacqueline Tedaldi, and Roya Høvsgaard, "Human Rights Impact Assessment Guidance and Toolbox." Danish Institute for Human Rights, 2016, https://www.socialimpactassessment.com/documents/hria_guidance_and_toolbox_final_jan2016.pdf

29 Article 29 Data Protection Working Party, "Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is 'Likely to Result in a High Risk' for the Purposes of Regulation 2016/679," WP 248 rev. 1, 2017, https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.

30 107th US Congress, *E-Government Act of 2002*.

EXISTING IMPACT ASSESSMENT PROCESSES

Fiscal Impact Assessment

In 2016, the City Council of Menlo Park needed to decide, as a **forum**, if it should permit the construction of a new mixed-use development, proposed by Sobato Corp. (the **actor**), near the center of town. They needed to know, prior to permitting (**time frame**), if the city could afford it, or if the development would **harm** residents by depriving them of vital city services: Would the new property and sales taxes generated by the development offset the costs to fire and police departments for securing its safety? Would the assumed population increase create a burden on the education system that it could not afford? How much would new infrastructure cost the city, beyond what the developers might pay for? Would the city have to issue debt to maintain its current standard of services to Menlo Park residents? Would this development be good for Menlo Park? To answer these questions, and to understand how the new development might **impact** the city's coffers, city planners commissioned a private company, BAE Urban Economics, to act as **assessors** and conduct a Fiscal Impact Assessment (FIA).[31] The FIA was **catalyzed** at the discretion of the City Council and was seen as having **legitimacy** based on the many other instances in which municipal governments looked to FIAs to inform their decision-making process.

By analyzing the city's finances for past years, and by analyzing changes in the finances of similar cities that had undertaken similar development projects, assessors were able to calculate the likely costs and revenues for city operations going forward—with and without the new development. The FIA process allowed a wide range of potential impacts to the people of Menlo Park—the quality of their children's education, the safety of their streets, the types of employment available to residents—to be made comparable by representing all these effects with a single metric, their impact to the city's budget. BAE compiled its analysis from existing fiscal statements (**method**) in a report, which the city gave **public access** to on its website.

With the FIA in hand, City Council members were able to engage in what is widely understood to be a "rational" form of governance. They weighed the pros against the cons and made an objective decision. While some FIA methods allow for more qualitative, contextual research and analysis, including **public participation**, the FIA process renders seemingly incomparable quality-of-life issues comparable by translating the issues into numbers, often collecting quantitative data from other places too, for the purposes of rational decision-making. Should the City Council make a "wrong" decision on behalf of Menlo Park's citizens, their only form of **redress** is at the ballot box, in the next election.

31 BAE Urban Economics, "Connect Menlo Fiscal Impact Analysis," City of Menlo Park Website, 2016, accessed March 22, 2021, https://www.menlopark.org/DocumentCenter/View/12112/Att-J_FIA.

THE CONSTITUTIVE COMPONENTS OF IMPACT ASSESSMENT

To build a framework for determining whether any proposed algorithmic impact assessment process is sufficiently complete to achieve accountability, we began with the five impact assessment processes listed in the previous section. We analyzed these impact assessment processes through historical examination of primary and secondary texts from their domains, examples of reporting documents, and examination of legislation and regulatory documents. From this analysis, we developed a schema that is common across all impact assessment regimes and can be used as an orienting principle to develop an AIA regime.

We propose that an ongoing process of consensus on arrangement of these **10 constitutive components** is the foundation for establishing accountability within any given impact assessment regime. (Please refer to table on page 15, and the expanded table on page 50). Understanding these 10 components, and how they can succeed and fail in establishing accountability, provides a clear means for evaluating proposed and existing AIAs. In describing "failure modes" associated with these components in the subsections below, our intent is to point to the structural features of organizing these components that can jeopardize the goal of protecting against harms to people, communities, and society.

It is important to note, however, that impact assessment regimes do not begin with laying out clear definitions of these components. Rather, they develop over time; impact assessment regimes emerge and evolve from a mix of legislation, regulatory rulemaking, litigation, public input, and scholarship. The common (but not universal) path for impact assessment regimes is that a rulemaking body (legislature or regulatory agency) creates a mandate and a general framework for conducting impact assessments. After this initial mandate, a range of experts and stakeholders work towards a consensus over the meaning and bounds of "impact" in that domain. As impact assessments are completed, a range of stakeholders—civil society advocates, legal experts, critical scholars, journalists, labor unions, industry groups, among others—will leverage whatever avenues are available—courtrooms, public opinion, critical research—to challenge the specific methods of assessing impacts and their relationship with actual harms. As precedents are established, standards around what constitutes an adequate account of impacts becomes stabilized. This stability is never a given; rather it is an ongoing practical accomplishment. Therefore, the following subsections describe each component, by illustrating the various ways they might be stabilized, and the failure modes are most likely to derail the process.

SOURCES OF LEGITIMACY

Every impact assessment process has a **source of legitimacy** that establishes the validity and continuity of the process. In most cases, the **source of legitimacy** is the combination of an institutional body (often governmental) and a definitional document (such as legislation and/or a regulatory mandate). Such documents often specify features of the other constituent components but need not lay out all the detail of the accountability regime. For example, NEPA (and subsequent related legislation) is the source of legitimacy for EIAs. This legitimacy, however, not only comes from the details of the legislation, but from the authority granted to the EPA by Congress to enforce regulations. However, legislation and institutional bodies by themselves do not produce an accountability regime. They instantiate a much larger recursive process of democratic governance through a regulatory state where various stakeholders legitimize the regime by actively participating in, resisting, and enacting it through building expert consensus and litigation.

Constitutive Component	Description
Sources of Legitimacy	Impact Assessments (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 with the power to impute responsibility to actors.
Actors and Forum	IAs are rooted in establishing an accountability relationship between *actors* who 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 its potential to enact accountability. Public access is essential to achieving transparency in the accountability relationship between actors and the forum.
Public Consultation	While public access governs transparency, *public consultation* creates conditions for solicitation of feedback from the broadest possible set of stakeholders in a system. Such consultations are resources to expand the list of impacts assessed or to shape the design of a system. Who constitutes this public and how are they consulted 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 for 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.

Other sources of legitimacy leave the specification of components open-ended. PIAs, for instance, get their legitimacy from a set of Fair Information Practice Principles (guidelines laid out by the Federal Trade Commission in the 1970s and codified into law in the Privacy Act of 1974[32]) but these principles do not explicitly describe how affected organizations should be held accountable. In a similar fashion, the Universal Declaration of Human Rights (UDHR) legitimizes HRIAs yet does not specify *how* HRIAs should be accomplished. Nothing under international law places responsibility for protecting or respecting human rights on corporations, nor are they required by any jurisdiction to conduct HRIAs or follow their recommendations. Importantly, while **sources of legitimacy** often define the basic parameters of an impact assessment regime (e.g., the *who* and the *when*), they often do not define every parameter (e.g., the *how*), leaving certain **constitutive components** to evolve organically over time.

Failure Modes for Sources of Legitimacy:

- *Vague Regulatory/Legal Articulations*: While legislation may need to leave room for interpretation of other constitutive components, being *too* vague may leave it ineffective. Historically, the tech industry has benefitted from its claims to self-regulate. Permitting self-regulation to continue unabated undermines the legitimacy of any impact assessment process.[33] Additionally, in an industry that is characterized by a complex technical stack involving multiple actors in the development of an algorithmic system, specifying the set of actors who are responsible for integrated components of the system is key to the legitimacy of the process.

- *Purpose Mismatch*: Different stakeholders may perceive an impact assessment process to serve divergent purposes. This difference may lead to disagreements about what the process is intended to do and to accomplish, thereby undermining its legitimacy. Impact assessments are *political*, empowering various stakeholders in relation to one another, and thus, influence key decisions. These politics often manifest in differences in rationales for why assessment is being done in the first place[34] in the pursuit of making a practical determination of whether to proceed with a project or not.[35] Making these intended purposes clear is crucial for appropriately bounding the expectations of interested parties.

32 Office of Privacy and Civil Liberties, "Privacy Act of 1974," *US Department of Justice*, https://www.justice.gov/opcl/privacy-act-1974; Federal Trade Commission, "Privacy Online: A Report to Congress," US Federal Trade Commission, 1998, https://www.ftc.gov/sites/default/files/documents/ reports/privacy-online-report-congress /priv-23a.pdf; Secretary's Advisory Committee on Automated Personal Data Systems, "Records, Computers, and the Rights of Citizens: Report," DHEW No. (OS) 73–94, US Department of Health, Education & Welfare, 1973, https://aspe.hhs.gov/report/records-computers-and-rights-citizens.

33 Edelman, Lauren B., 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, by Christine Parker and Vibeke Nielsen. Edward Elgar Publishing*. https://doi.org/10.4337/9780857938732.00011.; https://openscholarship.wustl.edu/law_lawreview/vol97/iss3/7

34 The form of rationality itself may be a point of conflict, as it may be an ecological rationality or an economic rationality. See: Robert V. Bartlett, "Rationality and the Logic of the National Environmental Policy Act," *Environmental Professional* 8, no. 2, (1986): 105–11.

35 Matthew Cashmore, Richard Gwilliam, Richard Morgan, Dick Cobb, and Alan Bond, "The Interminable Issue of Effectiveness: Substantive Purposes, Outcomes and Research Challenges in the Advancement of Environmental Impact Assessment Theory," *Impact Assessment and Project Appraisal* 22, no. 4 (2004): 295–310, https://doi.org/10.3152/147154604781765860.

- *Lack of Administrative Capacity to Conduct Impact Assessments*: The presence of legislation does not necessarily imply that impact assessments will be conducted. In the absence of administrative as well as financial resources, an impact assessment may simply remain a tenet of best practices.

- *Absence of Well-recognized Community/ Social Norms*: Creating impact assessments for highly controversial topics may simply not be able to establish legitimacy in the face of ongoing public debates regarding disagreements about foundational questions of values and expectations about whose interests matter. The absence of established norms around these values and expectations can often be used as defense by organizations in the face of adverse real-world consequences of their systems.

ACTORS AND FORUM

At its core, a **source of legitimacy** establishes a relationship between an **accountable actor** and an **accountability forum**. This relationship is most clear for EIAs, where the project developer—the energy company, transportation department, or Army Corps of Engineers—is the accountable actor who presents their project proposal and a statement of its expected environmental impacts (EIS) to the permitting agency with jurisdiction over the project. The permitting agency—the Bureau of Land Management, the EPA, or the state Department of Environmental Quality—acts as the accountability forum that can interrogate the proposed development, investigate the expected impacts and the reasoning behind those expectations, and can request alterations to minimize or mitigate expected impacts. The accountable actor can also face consequences from the forum in the form of a rejected or delayed permit, along with the forfeiture of the effort that went into the EIS and permit application.

However, the dynamics of this relationship may not always be as clear-cut. The forum can often be rather diffuse. For example, for FIAs, the accountable actor is the municipal official responsible for approving a development project, but the forum is all their constituents who may only be able to hold such officials accountable through electoral defeat or other negative public feedback. Similarly, PIAs are conducted by the government agency deploying an algorithmic system, however, there is no single forum that can exercise authority over the agency's actions. Rather, the agency may face applicable fines under other laws and regulations or reputational harm and civil penalties. The situation becomes even more complicated with HRIAs. A company not only makes *itself* accountable for the impacts of its business practices to human rights by commissioning an HRIA, but also acts as its own forum in deciding which impacts it chooses to address, and how. In such cases, as with PIAs, the public writ large may act as an alternative forum through censure, boycott, or other reputational harms. Crucially, many of the proposed aspects of algorithmic impact assessment assume this same conflation between actor and forum.

Failure Modes for Actors & Forum:

- *Actor/Forum Collapse*: There are many problems when actors and forums manifest within the same institution. While it is in theory possible for actor and forum to be different parties within one institution (e.g., ombudsman or independent counsel), the actor must be accountable to an external forum to achieve *robust* accountability.

- *A Toothless Forum*: Even if an accountability forum is external to the actor, it might not

have the necessary power to mandate change. The forum needs to be empowered by the force of law or persuasive social, political, and economic norms.

- *Legal Endogeneity*: Regulations sometimes require companies to demonstrate compliance, but then let them choose how, which can result in performative assessments wherein the forum abdicates to the actor its role in defining the parameters of an adequately robust assessment process.[36] This lends itself to a superficial checklist-style of compliance, or "ethics washing."[37]

CATALYZING EVENT

A **catalyzing event** triggers an impact assessment. Such events might be specified in law, for example, as NEPA specifies that an EIA is required in the US when proposed developments receive federal (or certain state-level) funding, or when such developments cross state lines. Other forms of impact assessment might be triggered on a more *ad hoc* basis, for example, an FIA is triggered when a municipal government decides, through deliberation, that one is necessary for evaluating whether to permit a proposed project. Along similar lines, a private company may elect to do an HRIA, either out of voluntary due diligence, or as a means of repairing its reputation following a public outcry, as was the case with Nike's HRIA following allegations of exploitative child labor throughout its global supply chain.[38] Impact assessment can also be anticipated within project development itself. This is particularly true for software development, where proper documentation throughout the design process can facilitate a future AIA.

Failure Modes for Catalyzing Events:

- *Exemptions within Impact Assessments*: A catalyzing event that exempts broad categories of development will have a limited effect on minimizing harms. If legislation leaves too many exceptions, actors can be expected to shift their activities to "game" the catalyst or dodge assessment altogether.

- *Inappropriate Theory of Change*: If catalyzing events are specified without knowledge of how a system might be changed, the findings of the assessment process might be moot. The timing of the catalyzing event must account for how and when a system can be altered. In the case of PIAs, for instance, catalysts can be at any point before system launch, which leads critics to worry that their results will come too late in the design process to effect change.

36 Lauren B. Edelman, and Shauhin A. Talesh, "To Comply or Not to Comply – That Isn't the Question: How Organizations Construct the Meaning of Compliance," in *Explaining Compliance*, by Christine Parker and Vibeke Nielsen, (Edward Elgar Publishing, 2011), https://doi.org/10.4337/9780857938732.00011.

37 Ben Wagner, "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?" in *Being Profiled*, edited by Emre Bayamlioglu, Irina Baralicu, Liisa Janseens, and Mireille Hildebrant, *Cogitas Ergo Sum: 10 Years of Profiling the European Citizen* (Amsterdam University Press, 2018), 84–89,https://doi.org/10.2307/j.ctvhrd092.18.

38 Nike, Inc., "Sustainable Innovation Is a Powerful Engine for Growth: FY14/15 Nike, Inc. Sustainable Business Report" Nike Inc., https://purpose-cms-production01.s3.amazonaws.com/wp-content/uploads/2018/05/14214951/NIKE_FY14-15_Sustainable_Business_Report.pdf.

EXISTING IMPACT ASSESSMENT PROCESSES

Environmental Impact Assessment

In 2014, Anadarko Petroleum Co. (the **actor**) opted to exercise their lease on US Bureau of Land Management (BLM) land by constructing dozens of coalbed methane gas wells across 1,840 acres of northeastern Wyoming.39 Because the proposed construction was on federal land, it **catalyzed** an Environmental Impact Assessment (EIA) as part of Anadarko's application for a permit that needed to be approved by the BLM (the **forum**),which demonstrated compliance with the National Environmental Protection Act (NEPA) and other environmental regulations that gave the EIA process its **legitimacy**. Anadarko hired Big Horn Environmental Consultants to act as **assessors**, conducting the EIA and preparing an Environmental Impact Statement (EIS) for BLM review as part of the permitting process.

To do so, Big Horn Environmental Consultants sent fieldworkers to the leased land and documented the current quality of air, soil, and water, the presence and location of endangered, threatened, and vulnerable species, and the presence of historic and prehistoric cultural materials that might be **harmed** by the proposed undertaking. With reference to several decades of scientific research on how the environment responds to disturbances from gas development, Big Horn Environmental Consultants analyzed the engineering and operating plans provided by Anadarko and compiled an EIS stating whether there would be **impacts** to a wide range of environmental resources. In the EIS, Big Horn Environmental Consultants graded impacts according to their severity, and recommended steps to mitigate those impacts where possible (the **method**). Where impacts could not be fully mitigated, permanent impacts to environmental resources were noted. Big Horn Environmental Consultants evaluated environmental impacts in comparison to a smaller, less impactful set of engineering plans Anadarko also provided, as well as in comparison to the likely effects on the environment if no construction were to take place (i.e., from natural processes like erosion or from other human activity in the area).

Upon receiving the EIS from Big Horn Environmental Consultants, the BLM evaluated the potential impacts on a **time frame** prior to deciding to issue a permit for Anadarko to begin construction. As part of that evaluation, the BLM had to balance the administrative priorities of other agencies involved in the permitting decision (e.g., Federal Energy Regulatory Commission, Environmental Protection Agency, Department of the Interior), the sometimes-competing definitions of impacts found in laws passed by Congress after NEPA (e.g., Clean Air Act, Clean Water Act, Endangered Species Act), as well as various agencies' interpretations of those acts. The BLM also gave **public access** to the EIS and opened a period of **public participation** during which anyone could comment on the proposed undertaking or the EIS. In issuing the permit, the BLM balanced the needs of the federal and state government to enable economic activity and domestic energy production goals against concerns for the sustainable use of natural resources and protection of nonrenewable resources.

39 Bureau of Land Management, *Environmental Assessment for Anadarko E&P Onshore LLC Kinney Divide Unit Epsilon 2 POD, WY-070-14-264* (Johnson County, WY: Bureau of Land Management, Buffalo Field Office, 2104), https://eplanning.blm.gov/public_projects/nepa/67845/84915/101624/KDUE2_EA.pdf.

TIME FRAME

When impact assessments are standardized through legislation (such as EIAs, DPIAs, and PIAs), they are often stipulated to be conducted within specific **time frames**. Most impact assessments are performed *ex ante* before a proposed project is undertaken and/or system is deployed. This is true of EIAs, FIAs, and DPIAs, though EIAs and DPIAs do often involve ongoing review of how actual consequences compare to expected impacts. FIAs are seldom examined after a project is approved.[40] Similarly, PIAs are usually conducted *ex ante* alongside system design. Unlike these assessments, HRIAs (and most other types of social impact analyses) are conducted *ex post*, as a forensic investigation to detect, remedy, or ameliorate human rights impacts caused by corporate activities. **Time frame** is thus both a matter of conducting the review before or after deployment, and of iteration and comparison

Failure Modes for Time Frame:

- *Premature Impact Assessments*: An assessment can be conducted too early, before important aspects of a system have been determined and/or implemented.
- *Retrospective Impact Assessments*: An *ex post* impact assessment is useful for learning lessons to apply in the future but does not address existing harms. While some HRIAs, for example, assess ongoing impacts, many take the form of after-action reports.
- *Sporadic Impact Assessments*: Impact assessments are not written in stone, and the potential impacts they anticipate (when conducted in the early phases of a project) may not be the same as the impacts that can be identified during later phases of a project. Additionally, assessments that speak to the scope and severity of impacts may prove to be over- or under-estimated once a project "goes live."

PUBLIC ACCESS

Every impact assessment process must specify its level of **public access**, which determines who has access to the impact statement reports, supporting evidence, and procedural elements. Without **public access** to this documentation, the **forum** is highly constrained, and its **source of legitimacy** relies heavily on managerial expertise. The broader the access to its impact statement, the stronger is an impact assessment's potential to enact changes in system design, deployment, and operation.

For EIAs, public disclosure of an environmental impact statement is mandated legislatively, coinciding with a mandatory period of public comment. For FIAs, fiscal impact reports are usually filed with the municipality as matters of public record, but local regulations vary. PIAs are public but their technical complexity often obscures more than it reveals to a lay public and thus, they have been subject to strong criticism. Or, in some cases in the US, a regulator has required a company to produce and file quasi-private PIA documents following a court settlement over privacy violations; the regulator holds it in reserve for potential future action, thus standing as a proxy for the public. Finally, DPIAs and HRIAs are only made public at the discretion of the company commissioning them. Without a strong commitment to make the

40 Robert W. Burchell, David Listokin, William R. Dolphin, Lawrence Q. Newton, and Susan J. Foxley, *Development Impact Assessment Handbook* (Washington, DC: Urban Land Institute, 1994), cited in: Edwards and Huddleston, 2009.

assessment accessible to the public at the outset, the company may withhold assessments that cast it in a negative light. Predictably, this raises serious concerns around the effectiveness of DPIAs and HRIAs.

Failure Modes for Public Access:

- *Secrecy/Inadequate Solicitation*: While there are many good reasons to keep elements of an impact assessment process private—trade secrets, privacy, intellectual property, and security—impact assessments serve as an important *public* record. If too many results are kept secret, the public cannot meaningfully protect their interests.

- *Opacities of Impact Assessments*: The language of technical system description, combined with the language of federal compliance and the potential length, complexity, and density of an impact assessment that incorporates multiple types of assessment data, can potentially enact a soft barrier to real public access to how a system would work in the real world.[41] For the lay public to truly be able to access assessment information requires ongoing work of translation.

PUBLIC CONSULTATION

Public consultation refers to the process of providing evidence and other input as an assessment is being conducted, and it is deeply shaped by an assessment's **time frame**. **Public access** is a precondition for **public consultation**. For *ex ante* impact assessments, the public, at times, can be consulted to include their concerns about or help reimagine a project. An example is how the siting of individual wind turbines becomes contingent on public concerns around visual intrusion to the landscape. Public consultation is required for EIAs, in the form of open comment solicitations, as well as targeted consultation with specific constituencies. For example, First Nation tribal authorities are specifically engaged in assessing the impact of a project on culturally significant land and other resources. Additionally, in most cases the forum is also obligated to solicit public comments on the merits of the impact statement and respond in good faith to public opinion.

Here the question of what constitutes a "public" is crucial. As various "publics" vie for influence over a project, struggles often emerge between social groups such as landowners, environmental advocacy organizations, hunting enthusiasts, tribal organizations, chambers of commerce, etc., for EIAs. For other *ex ante* forms of impact assessment, public consultation can turn into a hollow requirement, as with PIAs and DPIAs that mandate it without specifying its goals beyond mere notification. At times, public consultation can take the form of evidence gathered to complete the IA, such as when FIAs engage in public stakeholder interviews to determine the likely fiscal impacts of a development project.[42] Similarly, HRIAs engage the public in rightsholder interviews to determine how their rights have been affected as a key evidence-gathering step in conducting them.

41 Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (2016), https://doi.org/10.1177/2053951715622512.

42 Kotval and Mullin, 2006.

Failure Modes for Public Consultation:

- *Exploitative Consultation*: Public consultation in an impact assessment process can strengthen its rigor and even improve the design of a project. However, public consultation requires work on the part of participants. To ensure that impact assessments do not become exploitative, this time and effort should be recognized, and in some cases, compensated.[43]

- *Perfunctory Consultation*: Just because public consultation is mandated as part of an impact assessment, it does not mean that it will have any effect on the process. Public consultation can be perfunctory when it is held out of obligation and without explicit requirements (or strong norms).[44]

- *Inaccessibility*: Engaging in public consultation takes effort, and some may not be able to do so without facing a personal cost. This is particularly true of vulnerable individuals and communities, who may face additional barriers to participation. Furthermore, not every community that should be part of the process is aware of the harms they could experience or the existence of a process for redress.

METHOD

Standardizing **methods** is a core challenge for impact assessment processes, particularly when they require utilizing expertise and metrics across domains. However, **methods** are not typically dictated by **sources of legitimacy**, and are left to develop organically through regulatory agency expertise, scholarship, and litigation. Many established forms of impact assessment have a roster of well-developed and standardized **methods** that can be applied to particular types of projects, as circumstances dictate.[45]

The differences between **methods**, even within a type of impact assessment, are beyond the scope of this report, but they have several common features: *First,* impact assessment **methods** strive to determine what the impacts of a project will be relative to a counterfactual world in which that project does not take place. *Second,* many forms of expertise are assembled to comprise any impact assessment. 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. *The more varied the types of methods employed in an assessment process, the wider the range of impacts that can be assessed,* but likewise the greater expense of resources will be demanded. *Third,* impact assessment mandates a method for assembling information in a format that makes it possible for a forum to render judgement. PIAs, for example,

43 Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano, "Participation Is Not a Design Fix for Machine Learning," in *Proceedings of the 37th International Conference on Machine Learning*, 7 (Vienna, Austria: 2020).

44 Participation exists on a continuum for tokenistic, performative types of participation to robust, substantive engagement, as outlined by Arnstein's Ladder [Sherry R. Arnstein, "A Ladder of Citizen Participation," *Journal of the American Planning Association* 85 no. 1 (2019): 12.] and articulated for data governance purposes in work conducted by the Ada Lovelace Institute (personal communication with authors, March 2021).

45 See https://iaia.org/best-practice.php for an in-depth selection of impact assessment methods.

compile, in a single document, how a service will ensure that private information is handled in accordance with each relevant regulation governing that information.[46]

Failure Modes for Methods:

- *Disciplinarily Narrow*: Sociotechnical systems require methods that can address their simultaneously technical and social dimensions. The absence of diversity in expertise may fail to capture the entire gamut of impacts. Overly technical assessments with no accounting for human experience are not useful, and vice versa.

- *Conceptually Narrow*: Algorithmic impacts arise from algorithmic systems' actual or potential effects on the world. Assessment methods that do not engage with the world—e.g., checklists or closed-ended questionnaires for developers—do not foster engagement with real-world effects or the assessment of novel harms.

- *Distance between Harms and Impacts*: Methods also account for the distance between harms and how those harms are measured as impacts. As methods are developed, they become standardized. However, new harms may exceed this standard set of impacts. Robust accountability calls for frameworks that align the impacts, and the methods for assessing those impacts, as closely as possible to harms.

ASSESSORS

Assessors are those individuals (distinct from either actors or forum) responsible for generating an impact assessment. Every aspect of an impact assessment is deeply connected with *who* conducts the assessment. As evident in the case of HRIAs, accountability can become severely limited when the accountable actor and the accountability forum are collapsed within the same organization. To resolve this, HRIAs typically use external consultants as **assessors**.

Consulting group Business for Social Responsibility (BSR)—the assessors commissioned by Facebook to study the role of apps in the Facebook ecosystem in the genocide in Myanmar—is a prominent example. Their independence, however, must navigate a thin line between satisfying their clients and maintaining their independence. Other impact assessments—particularly EIA and FIAs—use consultants as assessors, but these consultants are subject to scrutiny by truly independent forums. For PIAs and DPIAs, the assessors are internal to the private company developing a data technology product. However, DPIAs may be outsourced if a company is too small, and PIAs rely on a clear separation of responsibilities across several departments within a company.

Failure Modes for Assessors:

- *Inexpertise*: Less mature forms of impact assessment may not have developed the necessary expertise amongst assessors for assessing impacts.

- *Limited Access*: Robust impact assessment processes require assessors to have broad access to full design specifications. If assessors are unable to access proprietary

46 Privacy Office of the Office Information Technology, "Privacy Impact Assessment (PIA) Guide," US Securities and Exchange Commission.

information—about trade secrets such as chemical formulae, engineering schematics, et cetera—they must rely on estimates, proxies, and hypothetical models.

- *Incompleteness*: Assessors often contend with the challenge of delimiting a complete set of harms from the projects they assess. Absolute certainty that the full complement of harms has been rendered legible through their assessment remains forever elusive and relies on a never-ending chain of justification.[47] Assessors and forums should not prematurely and/or prescriptively foreclose upon what must be assessed to meet criteria for completeness—new criteria can and do arise over time.

- *Conflicts of Interest*: Even formally independent assessors can become dependent on a favorable reputation with industry or industry-friendly regulators that could soften their overall assessments. Conflicts of interest for assessors should be anticipated and mitigated by alternate funding for assessment work, pooling of resources, or other novel mechanisms for ensuring their independence.

IMPACTS

Impact assessment is the task of determining what will be evaluated as a potential **impact**, what levels of such an impact are acceptable (and to whom), how such determination is made through gathering of necessary information, and finally, how the risk of an impact can be offset through financial compensation or other forms of redress. While impacts will look different in every domain, most assessments define them as *counterfactuals*, or measurable changes from a world without the project (or with other alternatives to the project). For example, an EIA assesses impacts to a water resource by estimating the level of pollutants likely to be present when a project is implemented as compared to their levels otherwise.[48] Similarly, HRIAs evaluate impact to specific human rights as abstract conditions, relative to the previous conditions in a particular jurisdiction, irrespective of how harms are experienced on the ground.[49] Along these lines, FIA assesses the future fiscal situation of a municipality after a development is completed, compared to what it would have been if alternatives to that development had taken place.[50]

Failure Modes for Impacts:

- *Limits of Commensuration*: Impact assessments are a process of developing a common metric of impacts that classifies, standardizes, and most importantly, makes sense of diverse possible harms. Commensuration,

47 Metcalf et al., "Algorithmic Impact Assessments and Accountability."

48 Richard K. Morgan, "Environmental impact assessment: the state of the art," *Impact Assessment and Project Appraisal* 30, no. 1 (March 2012): 5–14, https://doi.org/10.1080/14615517.2012.661557.

49 Deanna Kemp and Frank Vanclay, "Human rights and impact assessment: clarifying the connections in practice," *Impact Assessment and Project Appraisal* 31, no. 2 (June 2013): 86–96, https://doi.org/10.1080/14615517.2013.782978.

50 See, for example, Robert W. Burchell, David Listokin, and William R. Dolphin, *The New Practitioner's Guide to Fiscal Impact Analysis*, (New Brunswick, NJ: Center for Urban Policy Research, 1985); and Zenia Kotval and John Mullin, *Fiscal Impact Analysis: Methods, Cases, and Intellectual Debate,* Technical Report, Lincoln Institute of Land Policy, 2006.

the process of ensuring that terminology and metrics are adequately aligned among participants, is necessary to make impact assessments possible, but will inevitably leave some harms unaccounted for.

- *Limits of Mitigation*: Impacts are often not measured in a way that supports mitigation of harms. That is, knowing the negative impacts of a proposed system does not necessarily yield consensus over possible solutions to mitigate the projected harms.

- *Limits of a Counterfactual world*: Comparing the impact of a project with respect to a counterfactual world where the project does not take place inevitably requires making assumptions about what this counterfactual world would be like. This can make it harder to make arguments for *not* implementing a project in the face of projected harms, because they need to be balanced against the projected benefits of the project. Thinking through the uncertainty of an alternative is often hard in the face of the certainty offered by a project.

HARMS AND REDRESS

The **impacts** that are assessed by an impact assessment process are not synonymous with the **harms** addressed by that process, or how these harms are **redressed**. While FIAs assess impacts to municipal coffers, these are at least one degree removed from the harms produced. A negative fiscal impact can potentially result in declines in city services—fire, police, education, and health departments—which harm residents. While these harms are the implicit background for FIAs, the FIA process has little to do with how such harms are to be redressed, should they arise. The FIA only informs decision-making around a proposed development project, not the practical consequences of the decision itself.

Similarly, EIAs assess impacts to environmental resources, but the implicit harms that arise from those impacts are environmental degradation, negative health outcomes from pollution, intangible qualities like despoliation of landscape and viewshed, extinction, wildlife population decimation, agricultural yields (including forestry and animal husbandry), destruction of cultural properties and areas of spiritual significance. The EIA process is intended to address the likelihood of these harms through a well-established scientific research agenda that links particular impacts to specific harms. Therefore, the EIA process places emphasis on mitigation—requirements that funds be set aside to restore environmental resources to their prior state following a development—in addition to the minimization of impacts through the consideration of alternative development plans that result in lesser impacts.

If an EIA process is adequate, then there should be few, if any, unanticipated harms, and too many unanticipated harms would signal an inadequate assessment or a project that diverged from its original proposal, thus giving standing for those harmed to seek redress. For example, this has played out recently as the Dakota Access Pipeline project was halted amid courthouse findings that the EIA was inadequate.[51] While costly, litigation has over time

51 Scott K. Johnson, "Amid Oil- and Gas-Pipeline Halts, Dakota Access Operator Ignores Court," *Ars Technica*, July 8, 2020, https://arstechnica.com/science/2020/07/keystone-xl-dakota-access-atlantic-coast-pipelines-all-hit-snags/; Hiroko Tabuchi and Brad Plumer, "Is This the End of New Pipelines?" *The New York Times*, July, 2020, https://www.nytimes.com/2020/07/08/climate/dakota-access-keystone-atlantic-pipelines.html.

refined the bounds of what constitutes an adequate EIA, and the responsibilities of specific actors.[52]

The distance between impacts and harms can be even starker for HRIAs. For example, the HRIA[53] commissioned by Facebook to study the human rights impacts around violence and disinformation in Myanmar catalyzed by the refugee crisis, neither used the word "refugee" or common synonyms, nor directly acknowledged or recognized the ensuing genocide [see Human Rights Impact Assessment on page 27]. Instead, "impacts" to rights holders were described as harms to abstract rights such as security, privacy, and standard of living, which is a common way to address the constructed nature of impacts. Since the human rights framework in international law only recognizes nation-states, any harms to individuals found through this impact assessment could only be redressed through local judicial proceedings. Thus, actions taken by a company to account for and redress human rights impacts they have caused or contributed to remains strictly voluntary.[54] For PIAs and DPIAs, harms and redress are much more closely linked. Both impact assessment processes require accountable actors to document mitigation strategies for potential harms.

Failure Modes for Harms & Redress:

- *Unassessed Harms*: Given that harms are only assessable once they are rendered as impacts, an impact assessment process that does not adequately consider a sufficient range of harms within its scope of impacts, or inadequately exhausts the scope of harms that are rendered as impacts, will fail to address those harms.

- *Lack of Feedback*: When harms are unassessed, the affected parties may have no way of communicating that such harms exist and should be included in future assessments. For the impact assessment process to maintain its legitimacy and effectiveness, lines of communication must remain open between those affected by a project and those who design the assessment process for such projects.

52 Reliance on the courts to empower all voices excluded from or harmed by an impact assessment process, however, is not a panacea. The US courts have, until very recently (Hiroko Tabuchi and Brad Plumer. "Is This the End of New Pipelines?" *The New York Times*, July 8, 2020, https://www.nytimes.com/2020/07/08/climate/dakota-access-keystone-atlantic-pipelines.html), not been reliable guarantors of the equal protection of minority—particularly Black, Brown, and Indigenous—communities throughout the NEPA process. Pointing out that government agencies generally "have done a poor job protecting people of color from the ravages of pollution and industrial encroachment," (Robert D. Bullard, "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, 1999)) scholars of environmental racism argue that "the siting of unwanted facilities in neighborhoods where people of color live must not be seen as a *failure* of environmental law, but as a *success* of environmental law." (Luke W. Cole, "Remedies for Environmental Racism: A View from the Field," *Michigan Law Review* 90, no. 7 [June 1992]: 1991, https://doi.org/10.2307/1289740.) This is borne out by analyses of EIAs that fail to assess adverse impacts to communities located closest to proposed sites for dangerous facilities and also fail to adequately consider alternate sites—leaving sites near minority communities as the only "viable" locations for such facilities (*Ibid.).*

53 BSR, *Human Rights Impact Assessment: Facebook in Myanmar*, Technical Report, 2018, https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf.

54 Mark Latonero and Aaina Agarwal, "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar," Carr Center for Human Rights Policy, Harvard Kennedy School, 2021.

EXISTING IMPACT ASSESSMENT PROCESSES

Human Rights Impact Assessment

In 2018, Facebook (the **actor**) faced increasing international pressure[55] regarding its role in violent conflict in Myanmar, where over half a million Rohingya refugees were forced to flee to Bangladesh.[56] After that **catalyzing event**, Facebook hired an external consulting firm, Business for Social Responsibility (BSR, the **assessor**), to undertake a Human Rights Impact Assessment (HRIA). BSR was tasked with assessing the "actual impacts" to rights holders in Myanmar resulting from Facebook's actions. BSR's **methods**, as well as their **source of legitimacy**, drew from the UN Guiding Principles on Business and Human Rights[57] (UNGPs): Officials from BSR conducted desk research, such as document review, in addition to research in the field, including visits to Myanmar where they interviewed roughly 60 potentially affected rights holders and stakeholders, and also interviewed Facebook employees.

While actors and assessors are not mandated by any statute to give **public access** to HRIA reports, in this instance they did make public the resulting document (likewise, there is no mandated **public participation** component of the HRIA process). BSR reported that Facebook's actions had affected rights holders in the areas of security, privacy, freedom of expression, children's rights, nondiscrimination, access to culture, and standard of living. One risked impact on the human right to security, for example, was described as: "Accounts being used to spread hate speech, incite violence, or coordinate harm may not be identified and removed."[58] BSR also made several recommendations in their report in the areas of governance, community standards enforcement, engagement, trust and transparency, systemwide change, and risk mitigation. In the area of governance, BSR recommended, for example, the creation of a stand-alone human rights policy, and that Facebook engage in HRIAs in other high-risk markets.

However, the range of harms assessed in this solicited audit (which lacked any empowered **forum** or mandated **redress**) notably avoided some significant categories of harm: Despite many of the Rohingya being displaced to the largest refugee camp in the world[59], the report does not make use of the term "refugee" or any of its synonyms. It instead uses the term "rights holders" (a common term in human rights literature) as a generic category of person which does not name the specific type of harm that is at stake in this event. Further, the **time frame** of HRIAs creates a double-edged sword: assessment is conducted after a catalyzing event, and thus is both reactive to, yet cannot prevent, that event.[60] In response to the challenge of securing public trust in the face of these impacts, Facebook established their Oversight Board in 2020, which Mark Zuckerberg has often euphemized as the Supreme Court of Facebook, to independently address contentious and high-stakes moderation policy decisions.

55 Kevin Roose, "Forget Washington. Facebook's Problems Abroad Are Far More Disturbing," *The New York Times*, October 29, 2017, www.nytimes.com/2017/10/29/business/facebook-misinformation-abroad.html.

56 Libby Hogan and Michael Safi, "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.

57 United Nations Human Rights Office of the High Commissioner, "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework," New York and Geneva: United Nations, 2011, https://www.ohchr.org/Documents/Publications/ GuidingPrinciplesBusinessHR_EN.pdf.

58 BSR, *Human Rights Impact Assessment*.

59 World Food Program, "Rohingya Crisis: A Firsthand Look Into the World's Largest Refugee Camp," *World Food Program USA* (blog), 2020, accessed March 22, 2021, https://www.wfpusa.org/articles/rohingya-crisis-a-firsthand-look-into-the-worlds-largest-refugee-camp/.

60 Mark Latonero and Aaina Agarwal. 2021. "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar." Carr Center for Human Rights Policy Harvard Kennedy School.

TOWARD ALGORITHMIC IMPACT ASSESSMENTS

While we have found the **10 constitutive components** across all major impact assessments, no impact assessment regime emerges fully formed, and some constitutive components are more deliberately chosen or explicitly specified than others. The task for proponents of *algorithmic* impact assessment is to determine what configuration of these constitutive components would effectively govern algorithmic systems. As we detail below, there are multiple proposed and existing regulations that invoke "algorithmic impact assessment" or very similar mechanisms. However, they vary widely on how to assemble the constitutive components, how accountability relationships are stabilized, and how robust the assessment practice is expected to be. Many of the necessary components of AIAs already exist in some form; what is needed is clear decisions around how to assemble them. The striking feature of these AIA building blocks is the divergent (and partial) vision of how to assemble these constitutive components into a coherent governance mechanism.

In this section, we discuss existing and proposed models of AIAs in the context of the 10 constitutive components to identify the gaps that remain in constructing AIAs as an effective accountability regime. We then discuss algorithmic audits that have been crucial for demonstrating how AI systems cause harm. We will also explore internal technical audit and governance mechanisms that, while being inadequate for fulfilling the goal of robust accountability on their own, nevertheless model many of the techniques that are necessary for future AIAs. Finally, we describe the challenges of assembling the necessary expertise for AIAs.

Our goal in this analysis is not to critique any particular proposal or component as inadequate, but rather to point to the task ahead: assembling a consensus governance regime capable of capturing the broadest range of algorithmic harms and rendering them as "impacts" that institutions can act upon.

EXISTING & PROPOSED AIA REGULATIONS

There are already multiple proposals and existing regulations that make use of the term "algorithmic impact assessment." While all have merits, none share any consensus about how to arrange the **constitutive components** of AIAs. Evaluating each of these through the lens of the components reveals which critical decisions are yet to be made. Here we look at three cases: first, from proposals to regulate procurement of AI systems by public agencies; second, from an AIA currently in use in Canada; and third, one that has been proposed in the US Congress.

In one of the first discussions of AIAs, Andrew Selbst outlines the potential use of impact assessment methods for public agencies that procure automated decisions systems.[61] He lays out the importance of a strong regulatory requirement for AIAs (**source of legitimacy** and **catalyzing event**), the importance of **public consultation,** judicial review, and the consideration of alternatives.[62] He also emphasizes the need for an explicit focus on racial impacts.[63] While his focus is largely on algorithmic systems used in criminal justice contexts, Selbst notes a critically important aspect of impact assessment practices in general: that an obligation to conduct assessments is also an incentive to build the capacity to

61 Selbst, 2017.

62 *Ibid.*

63 Jessica Erickson, "Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System," *Washington Law Review* 89, no. 4 (2014): 1444–45.

understand and reflect upon what these systems actually do and whose lives are affected. Software procurement in government agencies is notoriously opaque and clunky, with the result that governments may not understand the complex predictive services that apply to all their constituents. By requiring an agency to account to the public how a system works, what it is intended to do, how the system will be governed, and what limitations the system may have can force at least a portion of the algorithmic economy to address widespread challenges of algorithmic explainability and transparency.

While Selbst lays out how impact assessment and accountability intersect in algorithmic contexts, AI Now's 2018 report proposes a fleshed-out framework for AIAs in public agencies.[64] Algorithmic systems present challenges for traditional governance instruments. While appearing similar to software systems regularly handled by procurement oversight authorities, they function differently, and might process data in unobservable "black-boxed" ways. AI Now's proposal recommends the New York City government as the **source of legitimacy** for adapting the procurement process to be a **catalyzing event**, which triggers an impact assessment process with a strong emphasis on **public access** and **public consultation**. Along these lines, the office of New York City's Algorithms Management and Policy Officer, in charge of designing and implementing a framework "to help agencies identify, prioritize, and assess algorithmic tools and systems that support agency decision-making,"[65] produced an Algorithmic Tool Directory in 2020. This directory identifies a set of algorithmic tools already in use by city agencies and is available for **public access**.[66] Similar efforts for transparency have been introduced at the municipal level in other major cities of the world, such as the accessible register of algorithms in use in public service agencies in Helsinki and Amsterdam.[67]

AIA requirements recently implemented by Canada's Treasury Board reflect aspects of AI Now's proposal. The Canadian Treasury Board oversees government spending and guides other agencies through procurement decisions, including procurement of algorithmic systems. Their AIA guidelines mandate that any government agency using such systems, or any vendor using such systems to serve a government agency, complete an algorithmic impact assessment, "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."[68] The actual form taken by the AIA is an electronic survey that is meant to help agencies

64 Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," *AI Now Institute*, 2018https://ainowinstitute.org/aiareport2018.pdf.

65 City of New York Office of the Mayor, *Establishing an Algorithms Management and Policy Officer*, Executive Order No. 50, 2019, https://www1.nyc.gov/assets/home/downloads/pdf/executive-orders/2019/eo-50.pdf.

66 Jeff Thamkittikasem, "Implementing Executive Order 50 (2019), Summary of Agency Compliance Reporting," City of New York Office of the Mayor, Algorithms Management and Policy Officer, 2020, https://www1.nyc.gov/assets/ampo/downloads/pdf/AMPO-CY-2020-Agency-Compliance-Reporting.pdf.

67 Khari Johnson, "Amsterdam and Helsinki Launch Algorithm Registries to Bring Transparency to Public Deployments of AI," *VentureBeat*, September 28, 2020, https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/.

68 Treasury Board of Canada Secretariat, "Directive on Automated Decision-Making," 2019 https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

EXISTING IMPACT ASSESSMENT PROCESSES

Data Protection Impact Assessment

In April 2020, amidst the COVID-19 global pandemic, the German Public Health Authority announced its plans to develop a contact-tracing mobile phone app.[69] Contact tracing enables epidemiologists to track who may have been exposed to the virus when a case has been diagnosed, and thereby act quickly to notify people who need to be tested and/or quarantined to prevent further spread. The German government's proposed app would use low-energy Bluetooth signals to determine proximity to other phones with the same app for which the owner has voluntarily affirmed a positive COVID-19 test result.[70]

The German Public Health Authority determined that this new project, called Corona Warn, would process individual data in a way that was likely to result in a high risk to "the rights and freedoms of natural persons," as determined by the EU Data Protection Directive Article 29. This determination was a **catalyst** for the public health authority to conduct a Data Protection Impact Assessment (DPIA).[71] The **time frame** for the assessment is specified as beginning before data is processed, and conducted in an ongoing manner. The **theory of change** requires that **assessors**, or "data controllers," think through their data management processes as they design the system to find and mitigate privacy risks. Assessment must also include **redress**, or steps to address the risks, including safeguards, security measures, and mechanisms to ensure the protection of personal data and demonstrate compliance with the EU's General Data Protection Regulation, the regulatory framework which also acts as the DPIA **source of legitimacy**.

Per the Article 29 Advisory Board,[72] **methods** for carrying out a DPIA may vary, but the criteria are consistent: **Assessors** must describe the data this system had to collect, why this data was necessary for the task the app had to perform, as well as modes for data processing-management risk mitigation. Part of this methodology must include consultation with data subjects, as the controller is required to seek the views of data subjects or their representatives where appropriate" (Article 35(9)). **Impacts**, as exemplified in the Corona Warn DPIA, are conceived as potential risks to the rights and freedoms of natural persons arising from attackers whose access to sensitive data is risked by the app's collection. Potential attackers listed in the DPIA include business interests, hackers, and government intelligence. Risks are also conceived as unlawful, unauthorized, or nontransparent processing or storage of data. **Harms** are conceived as damages to the goals of data protection, including damages to data minimization, confidentiality, integrity, availability, authenticity, resilience, ability to intervene, and transparency, among others. These are also considered to have downstream damage effects. The **public access** component of DPIAs is the requirement that resulting documentation be produced when asked by a local data protection authority. Ultimately, the **accountability forum** is the country's Data Protection Commission, which can bring consequences to bear on developers, including administrative fines as well as inspection and document seizure powers.

69 Rob Schmitz, "In Germany, High Hopes for New COVID-19 Contact Tracing App That Protects Privacy," *NPR*, April 2, 2020, https://www.npr.org/sections/coronavirus-live-updates/2020/04/02/825860406/in-germany-high-hopes-for-new-covid-19-contact-tracing-app-that-protects-privacy.

70 The Germany Public Health Authority altered the app's data-governance approach after public outcry, including the publication of an interest group's DPIA (Kristen Bock, Christian R. Kuhne, Rainer Muhlhoff, Meto Ost, Jorg Poole, and Rainer Rehak. "Data Protection Impact Assessment for the Corona App," Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FIfF) e.V., 2020, https://www.fiff.de/dsfa-corona) and a critical open letter from scientists and scholars ("JointStatement on Contact Tracing." 2020. https://main.sec.uni-hannover.de/JointStatement.pdf).

71 Article 29 Data Protection Working Party, "Guidelines on Data Protection Impact Assessment (DPIA)."

72 *Ibid.*

"evaluate the impact of automated decision-support systems including ethical and legal issues."[73] Questions include, "Are the impacts resulting from the decision reversible?"; "Is the project subject to extensive public scrutiny (e.g., 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?"[74] The survey instrument scores the answers provided to produce a risk score.[75]

Critics have pointed out[76] that such Yes/No-based self-reporting does not bring about insight into how these answers are decided, what metrics are used to define "impact" or "public scrutiny," or guarantee subject-matter expertise on such matters. While this system can enable an agency to create risk tiers to assist in choosing between vendors, it cannot fulfill the requirements of a **forum** for accountability, reducing its ability to protect vulnerable people. This rule has also come under scrutiny regarding its **sources of legitimacy** when Canada's Department of Defense determined that it did not need to submit an AIA for a hiring-diversity application because the system did not render the "final" decision on a candidate.[77]

These models for algorithmic governance in public agency procurement share constitutive components most similar to FIAs and PIAs. The **catalyst** is initiation of a public procurement process, the **accountable actor** is the procuring agency (although relying heavily on the vendor for information about how the system works), the **accountability forum** is the democratic process (i.e., elections, public comments) and litigation, the theory of change relies upon public pressuring representatives for high standards, the **time frame** is *ex ante*, and the **access** to documentation is public. The type of **harm** that these AIAs most directly address is a lack of transparency in public institutions—they do not necessarily audit or prevent downstream, concrete effects such as racial bias in digital policing. The harm is conceived as damage to democratic self-governance by displacing explicable, human-driven, sociopolitical decisions with machinic, inexplicable decisions. By addressing the algorithmic transparency problem, it becomes *possible* for advocates to address those more concrete **harms** downstream via public pressure to

73 Michael Karlin, "The Government of Canada's Algorithmic Impact Assessment: Take Two." https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f; Michael Karlin, "Deploying AI Responsibly in Government," *Policy Options* (blog), February 6, 2018, https://policyoptions.irpp.org/magazines/february-2018/deploying-ai-responsibly-in-government/.

74 Government of Canada, "Canada-ca/Aia-Eia-Js," JSON, Government of Canada, 2019,https://github.com/canada-ca/aia-eia-js.

75 Government of Canada, "Algorithmic Impact Assessment – Évaluation de l'Incidence Algorithmique." Algorithmic Impact Assessment, June 3, 2020, https://canada-ca.github.io/aia-eia-js/.

76 Mathieu Lemay, "Understanding Canada's Algorithmic Impact Assessment Tool," *Toward Data Science* (blog), June 11, 2019, https://towardsdatascience.com/understanding-canadas-algorithmic-impact-assessment-tool-cd0d3c8cafab.

77 Tom Cardoso and Bill Curry, "National Defence Skirted Federal Rules in Using Artificial Intelligence, Privacy Commissioner Says," *The Globe and Mail*, February 7, 2021,https://www.theglobeandmail.com/canada/article-national-defence-skirted-federal-rules-in-using-artificial/.

block or rescind procurement or via litigation (e.g., disparate impact cases).

The 2019 Algorithmic Accountability Act proposed to empower US federal regulatory agencies to require AIAs in regulated domains (e.g., financial loans, real estate, medicine, etc.).[78] In contrast to the above models focusing on public agency procurement, the bill 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 it (and to retroactively conduct an AIA for all existing systems). The bill's sponsors attempted to ensure that the nondiscrimination standards for economic activities in regulated domains (e.g., financial loans, real estate, medicine, etc.) are also applied to algorithmic systems.[79] The public regulator's requirements would include an assessment but permit the entity to decide for themselves whether to make the resulting algorithmic impact assessment documentation public (though it would be discoverable in civil or criminal legal proceedings). Such discretion means the standard would lack teeth: without a forum in which that assessment can be examined or judged, there is no public transparency to bring about accountability relationship between the **actors and forums**. As a contrast with the procurement-oriented AIAs, the act's model establishes the companies building and selling algorithmic systems as the **accountable actor**, a regulatory agency (as a proxy for the public interest) as the accountability **forum,** and the theory of change relies upon the forum to represent the public interest. Notably, the Algorithmic Accountability Act does not indicate the degree to which the public would have **access** to the AIA documentation, whether in whole or in part. This model is most analogous to the PIA process that occurs in some large tech companies, most notably those that are under consent decrees with the US regulatory agencies following privacy violations and enforcement actions (PIAs are not universally used in the tech industry as a governance document). As of the release of this report, public reporting has indicated that a version of the Algorithmic Accountability Act is likely to be reintroduced in the current Congress, providing an opportunity for reconsideration of how accountability will be structured.[80]

Notably, the European approach appears to be evolving in a different direction: toward a general obligation for developers to record and maintain documentation about how systems were trained and designed, describing in detail how higher-risk systems operate, and attesting to compliance with EU regulations. The European Commission's reports have emphasized establishing an "ecosystem of trust" that will encourage EU citizens to participate in the data economy.[81] The European Commission recently released the first formal

78 Yvette D. Clarke, "H.R.2231—116th Congress (2019–2020): Algorithmic Accountability Act of 2019," 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231.

79 Cory Booker, "Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms," *Press Office of Sen. Cory Booker* (blog), April 10, 2019, https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms.

80 Issie Lapowsky and Emily Birnbaum. 2021. "Democrats Have Won the Senate. Here's What It Means for Tech." Protocol—The People, Power and Politics of Tech. January 6, 2021. https://www.protocol.com/democrats-georgia-senate-tech.

81 European Commission, "On Artificial Intelligence – A European Approach to Excellence and Trust." White Paper (Brussels: 2020), https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf; Panel for the Future of Science and Technology, "A Governance Framework for Algorithmic Accountability and Transparency," EU: European Parliamentary Research Service, 2019, https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

draft of its AI regulatory framework, known by the shorthand Artificial Intelligence Act.[82, 83]

The act establishes a three-tiered regulatory model: prohibited systems, high-risk systems that require additional third-party auditing and oversight, and presumed-safe systems that can self-attest to compliance with the act. Many of the headlines have focused on the prohibitions on certain use cases (mass biometric surveillance, manipulation and disinformation, discrimination, and social scoring) and the definitions of high-risk systems, such as safety components, systems used in an already regulated domain, and applications with risk of harming fundamental human rights. As an analysis by the civil society group European Digital Rights points out, this proposed regulation is centered on self-governance by developers and largely relies on their own attestation of compliance with their governance obligations.[84] The proposed auditing, reporting and certification regime resembles impact assessments in a variety of ways. It establishes an accountability relationship between **actors** (developers) and a **forum** (notified body); it creates a partial form of **public access** through reporting and attestation requirements on an *ex ante* **time frame**; and the power of the notified body to conduct a conformity audit power is likely to spawn a variety of **methods**.

As Selbst noted[85], even the bureaucratic requirement to retain technical data and explain design decisions in anticipation of such an assessment is likely to provide a significant incentive for developers to build the internal capacity to make more deliberate and safer decisions about algorithmic systems.

Ultimately, the EU proposal shares more in common with industrial safety rules than impact assessment, with a strong emphasis on bureaucratic standardization and few opportunities for **public consultation** and contestation over the values and societal purpose of these algorithmic systems or opportunities for **redress**. Additionally, the act mostly regulates algorithmic systems by market domain—financial applications are regulated by finance regulators, medical application are regulated by medical regulators, et cetera—which disperses expertise in auditing algorithmic systems and public watchdog efforts across many different agencies. While this rule would provide a significant step forward in global algorithmic governance, there is reason to be concerned that the **assessors** and **methods** would be too distant from the lived experience of algorithmic harms.

Comparing these AIA models through the lens of **constitutive components,** it becomes clear

82 Council of Europe, and European Parliament. 2021. "Regulation on European Approach for Artificial Intelligence Laying Down a Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

83 As of the publication of this report, the Act is still in an early stage of the legislative process and is likely to undergo significant amendment as it is taken up by the European Parliament. The version discussed here is the first publicly available draft, released in April, 2021.

84 Chander, Sarah, and Ella Jakubowska. 2021. "EU's AI Law Needs Major Changes to Prevent Discrimination and Mass Surveillance." European Digital Rights (EDRi). https://edri.org/our-work/eus-ai-law-needs-major-changes-to-prevent-discrimination-and-mass-surveillance/.

85 Andrew Selbst, "Disparate Impact and Big-Data Policing."

EXISTING IMPACT ASSESSMENT PROCESSES

Privacy Impact Assessment

In 2013 a United States federal agency involved in issuing travel documents, such as visas and passports, decided to design a new data-driven program to help flag potential terrorism suspects in the millions of applications they receive every year. Their new system would use facial recognition technology to compare photos of people applying for travel documents, against federally collected images in databases maintained by counter-terrorism agencies. As are all federal agencies, they were obligated, per the E-Government Act of 2002, to evaluate the potential privacy impacts of their new system. For this evaluation, they would need to conduct a Privacy Impact Assessment (PIA). The **catalyst** for conducting the PIA was twofold: first, the design of a new system, and second, the fact it collected personally identifiable information (PII). The **assessor**, or person conducting the PIA, was the agency's Chief Information Coordinator.

The **method** the assessor used to conduct the PIA was to catalogue several attributes of the system, including where and how data was sourced, used, and shared; why that data was necessary for the goals of the agency; how these practices adhered to existing regulatory and policy mandates; the privacy risks engendered by these practices, and how those risks would be mitigated. The **time frame** in which the PIA was conducted was in tandem with the development of the system: Developers needed to think about how the systems they were building might affect the privacy of individuals, and further, how such impacts might create risks, down the line, for the agency itself. This time frame was key for the **theory of change** underpinning the PIA: Designers of the PIA process intended for the completion of the document to inculcate privacy awareness into developers who would, hopefully, build privacy-aware values into the system as they assessed it.[86]

The resulting report detailed that all practices complied with pre-established norms for managing data, in particular Title III of the aforementioned E-Government Act, the Federal Information Security Management Act (FISMA), as well as information assurance standards set by the National Institute of Standards and Technology (NIST). These norms and regulations made up the **source of legitimacy** for the PIA process: Thousands of experts, regulators, and legal scholars had worked together over several years to create and set these standards. Implementing these norms also formed the agency's approach to **redress** in the face of **harms**, or ways that they addressed and mitigated the risks that their data collection might have for individuals.

Lastly, the agency posted their PIA to their website as a PDF. Making this document **public** laid bare the decisions that were made about the system and constituted a type of **forum for accountability**. This transparency threatened punitive damages to the agency if they did not do the PIA correctly, if they had been found to have provided false information, or if they had failed to address dangers presented to individuals. Potential **impacts** to the agency included financial loss from fines, loss of public trust and confidence, loss of electoral support, cancelation of a project, penalties resulting from the infringement of laws or regulations leading to judicial proceedings, and/or the imposition of new controls in response to public concerns about the project, among others.[87]

86 Kenneth A. Bamberger and Deirdre K. Mulligan, "PIA Requirements and Privacy Decision-Making in US Government Agencies," in *Privacy Impact Assessment*, edited by David Wright and Paul De Hert (Dordrecht: Springer, 2012), 225–50, https://link.springer.com/chapter/10.1007/978-94-007-2543-0_10.

87 David Wright and Paul De Hert, "Introduction to Privacy Impact Assessment," in *Privacy Impact Assessment*, edited by David Wright and Paul De Hert, (Dordrecht: Springer, 2012), 3–32, https://link.springer.com/chapter/10.1007/978-94-007-2543-0_1.

that there is little agreement on how to structure accountability relationships. There is a lack of consensus on what an algorithmic harm is, how those harms should be rendered as impacts, and who should have the responsibility to force changes to the systems. Looking to the table of **constitutive components** on Appendix A, the challenge for advocates of AIAs moving forward is to articulate a coherent, common understanding of how to fill in these components, particularly for a **source of legitimacy** that conforms to the robust definition of accountability between an **actor** and a **forum**, and how to map **impacts** to **harms**.

ALGORITHMIC AUDITS

Prior to the current interest in AIAs, algorithmic systems have been subjected to a variety of internal and external "audits" to assess their effectiveness and potential consequences in the world. While audits alone are not generally suitable for robust accountability, they can nonetheless reveal effective techniques for assembling a number of the constituent components absent from current AIA proposals, and in some cases offer models for informing the public about the operation of such systems.

Technical auditing is a longstanding practice within and beyond[88] computing, and has become a core feature of the rapidly evolving field of algorithmic governance.[89] In computational contexts, auditing is the practice of comparing the functioning of a system against a benchmark and judging whether variance between the system and benchmark is within acceptable parameters and/or otherwise justified. That benchmark could be a technical description provided by the developer, an outcome prescribed in a contract, a procedure defined by a standards organization such as IEEE or ISO, commonly accepted best practices, or a regulatory mandate. Audits are performed by experts with the capacity to render such judgement, and with a degree of independence from the development process.[90] Across most domains, auditors can be described as: *third party*, someone outside of the audited organization with access to only the outputs of the system; *second party*, someone hired from outside the developing organization with access to the backend and outputs of the system; and *first party*, someone internal to the organization who is primarily conducting internal governance. Although this distinction does not yet circulate universally in algorithmic auditing, we make use of it here because it clarifies important features of auditing and illustrates the utility and limits of auditing for AIAs.[91]

External (Third- and Second-Party) Audits

Audits conducted by external, third-party **assessors** with no formal relationship to the developer have been a primary driver of the public attention to algorithmic harms and a motivating force for the development of internal governance mechanisms (also discussed below) that some tech companies have begun adopting. Notable examples include ProPublica's analysis of the Northpointe COMPAS

88 Michael Power, *The Audit Society: Rituals of Verification* (New York: Oxford University Press, 1997).

89 Ada Lovelace Institute "Examining the Black Box: Tools for Assessing Algorithmic Systems," Ada Lovelace Institute, 2020, https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

90 Even where the auditing is fully internal to a company, the auditor should not have been involved in the product

91 This schema is somewhat complicated by the rise of "collaborative audits" between developers and auditing entities who work together to delineate the scope and purpose of an audit. See, Mona Sloane, "The Algorithmic Auditing Trap," *OneZero* (blog), March 17, 2021, https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.

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.[92] In each of these cases, external experts analyzed algorithmic systems primarily through the *outputs* of deployed systems without access to the backend controls or models, which only happens after a system has already been deployed.[93] This is the core feature of adversarial third-party algorithmic audits: the assessor lacks access to the backend controls and design records of the system and therefore is limited to understanding the outputs of the opaque, black-boxed systems. Without access, an adversarial third party needs to rely on records of how the system operates in the field, from the epistemic position of observer rather than engineer.[94]

The diversity in algorithmic systems means different adversarial audits might be forced to rely on significantly different **methods**. For example, ProPublica's analysis of recidivism scores assigned by COMPAS in Broward County, Florida, relied upon what could be gleaned about the effects of the system from historical records without **public access** to the system.[95] In contrast, the Gender Shades audits used an artificially constructed "population" to compare the accuracy of multiple facial recognition services across demographic categories via their commercial APIs. This method known as a "sock puppet audit"[96] allowed the auditors to act as if end users.

Despite often having to innovate their methods in the absence of direct access to algorithmic systems, third-party audits create a **forum** out of publics writ large by bringing pressure to bear on the developers in the form of negative public attention.[97] But their externality is also a vulnerability: when the targets of these audits have engaged in rebuttals, their technical analyses have invoked knowledge of the

92 Buolamwini and Gebru, 2018; Eubanks, 2018.

93 Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," in *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, Vol. 22. (Seattle WA: 2014); Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris, "Detecting Price and Search Discrimination on the Internet," in *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI*, (Redmond, Washington: ACM Press, 2012), 79–84, https://doi.org/10.1145/2390231.2390245; Ben Green and Yiling Chen, "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*,FAT* '19 (New York, NY, USA: Association for Computing Machinery, 2019), 90–99, https://doi.org/10.1145/3287560.3287563.

94 Inioluwa Deborah Raji and Joy Buolamwini, "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 (New York, NY, USA: Association for Computing Machinery), 429–435, https://doi.org/10.1145/3306618.3314244; Joy Buolamwini, "Response: Racial and Gender Bias in Amazon Rekognition — Commercial AI System for Analyzing Faces," *Medium*, April 24, 2019, https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon- rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced.

95 Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, n.d., accessed March 22, 2021, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=6LHoUCqhSP02JHSsAi7mlAd73V6zJtgb.

96 Raji and Buolamwini, 2019; Sandvig and Langbort, 2014.

97 Joy Buolamwini, "Amazon Is Right: Thresholds and Legislation Matter, So Does Truth," *Medium* (blog), February 7, 2019, https://medium.com/@Joy.Buolamwini/amazon-is-right-thresholds-and- legislation-matter-so-does-truth-6cfdf6005c80.

systems' design parameters that an adversarial third-party auditor could not have had access to.[98] The reliance on such technical analyses in response to audits pointing out sociopolitical harms all too often fall into the trap of the *specification dilemma*: that is, prioritizing technical explanations for why a system might function as intended, while ignoring that accurate results might themselves be the source of harm. Inaccurate matches made by a facial recognition system may not be an algorithmic *harm*, but exclusionary consequences[99] that can all flow from misrecognition by a facial recognition technology certainly are algorithmic harms. A purely technical response to these harms is inadequate. In short, *third-party audits have illustrated how little the public knows about the actual functioning of the systems that render major decisions about our lives through algorithmic prediction and classification.*

As important as third-party audits have been for increasing public transparency into the operation of algorithmic systems, such audits cannot ever constitute robust algorithmic accountability. The third-party audit format is often motivated by the absence of a **forum** with the capacity to demand change from an **actor**, and relies on negative public attention to enact change, as fickle and lacking legal force as that may be.[100] This is manifested in the lack of a **catalyzing event** beyond the attention and commitment of the auditor, a mismatch between the **timeframe** of assessments and deployment, and the unofficial **source of legitimacy** that mostly consists of the professional reputation of the auditors and their ability to motivate public attention.

Perhaps the most important role of a **forum** is to be empowered by a **source of legitimacy** to set the conditions for rendering an informed judgement based on potentially very disparate sources of evidence. Consider as an example the Allegheny Family Screening Tool (AFST)—an algorithmic system used to assist child welfare call screening—arguably the most thoroughly audited algorithmic system in use by a public agency in the US. See the sidebar on page 46. The AFST was subject to procurement reviews and internal audits,[101] a solicited external

98 William Dietrich, Christina Mendoza, and Tim Brennan, "COMPAS Risk Scales: Demonstrating Accuracy, Equity and Predictive Parity," Northpointe Inc. Research Department, 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

99 Hill, "Wrongfully Accused by an Algorithm"; Moran, "Atlantic Plaza Towers Tenants Won a Halt to Facial Recognition"; and Brammer, "Trans Drivers Are Being Locked Out."

100 Indeed, Inioluwa Deborah Raji, a co-author of a Gender Shades audit, notes that the strategic purpose of third-party adversarial audits is to create pressure on companies to change their practices wholesale and on legislators to impose regulations covering algorithmic harms. See: "The Radical AI Podcast: With Deb Raji." June, 2020. The Radical AI Podcast. https://www.radicalai.org/e15-deb-raji; Raji, Inioluwa Deborah, 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, 429–35. AIES '19. New York, NY, USA: Association for Computing Machinery.* https://doi.org/10.1145/3306618.3314244.

101 Rhema Vaithianathan, Tim Maloney, Emily Putnam-Hornstein, and Nan Jiang, "Children in the Public Benefit System at Risk of Maltreatment: Identification via Predictive Modeling," *American Journal of Preventive Medicine 45, no. 3 (2013): 354–59,* https://doi.org/10.1016/j.amepre.2013.04.022; and Emily Putnam-Hornsteinand Barbara Needell, "Predictors of Child Protective Service Contact between Birth and Age Five: An Examination of California's 2002 Birth Cohort," *Children and Youth Services Review, Maltreatment of Infants and Toddlers, 33, no. 8 (2011): 1337–44,* https://doi.org/10.1016/j.childyouth.2011.04.006.

algorithmic fairness audit,[102] a second-party ethics audit,[103] and an adversarial third-party social science audit[104]. These audits produced significantly divergent and often conflicting results representing their respective **methods**, which at times rely on incommensurable frameworks. Robust accountability depends on collaboratively resolving *what we can know and how we should know it*. No matter the quality and diversity of auditing methods available, there remains the challenge of making those audits commensurable accounts of **impacts**, something that only a **legitimate,** empowered **forum** backed by consensus can do.

Indeed, it is this thoroughness paired with the widely divergent interpretations of the same system that highlights the limitations of audits without accountability relationships between an **actor** and an empowered **forum**. These disparate approaches for analyzing the consequences of algorithmic systems may be complementary but *they cannot contribute to a single, actionable interpretation without establishing institutional accountability through a consensus process for bounding impacts*. A third-party audit is limited in its ability to create a comprehensive picture of the consequences of a system and draw an actionable connection between design decisions and their **impacts**. Both third-party and second-party audits are further limited in forcing appropriate changes to the system insofar as they lack a formal **source of legitimacy**. The theory of change underlying third-party audits relies on fickle public attention forcing voluntary (but usually not structural) changes[105]; the result is a disempowered **forum** with an uncertain relation to an **actor**. The **time frame** for a third-party audit is capricious because it happens at any time after the outputs of the system become visible to the auditor, potentially long after harms have already been caused.

Second-party audits are likely closer in practice to much of the work that would be used to generate algorithmic impact statements, but likewise do not alone have an adequate answer for how to assemble all the **constitutive components**. Where a third-party audit is a **forum** without an **actor**, a second-party audit is an **actor** without a **forum,** unless a regulatory mandate is secured. Along the same lines, second-party audits can often proceed without **public consultation** or **public access** because the auditor is primarily responsive to the party that hired them, and in many cases may not be able to share proprietary information relevant to the public interest. Furthermore, without a consensus

102 Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan, "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," in *Conference on Fairness, Accountability and Transparency, 2018, 134–48,* http://proceedings.mlr.press/v81/chouldechova18a.html.

103 Tim Dare and Eileen Gambrill, "Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County," in Vaithianathan, 2017.

104 Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor (St. Martin's Press, 2018). In most contexts, Eubanks' work would not be identified as an "audit." An audit typically requires an established standard against which a system can be tested for divergence. However, the stakes with AIAs is that a broad range of harms must be accounted for, and thus analyses like Eubanks' would need to be made commensurate with technical audits in any sufficient AIA process. Therefore we use the term idiosyncratically. See: Josephine Seah, "Nose to Glass: Looking In to Get Beyond," ArXiv: 2011.13153 [Cs], December 2020,* http://arxiv.org/abs/2011.13153.

105 The authors of influential third-party audits readily acknowledge these limits. For example, data scientist Inioluwa Deborah Raji, co-author of the second Gender Shades audit and a number of internal auditing frameworks (discussed below), noted in an interview that the ultimate goal of adversarial third-party audits is to create pressure on technology companies and regulators that will lead to future robust regulatory obligations around algorithmic governance. See: "The Radical AI Podcast," *The Radical AI Podcast*, June 2020, https://www.radicalai.org/e15-deb-raji.

that bounds **impacts** such that algorithmic harms are accounted for, second-party auditors are constrained by the parameters set by those who contracted the audit.[106]

Internal (First-Party) Technical Audits & Governance Mechanisms

First-party audits are distinct from other forms of audits in that they are performed for the purpose of satisfying the developer's own concerns. Those concerns may be indexed to common elements of responsible AI practice, like transparency and fairness, which may be due entirely to magnanimous reasons or for utilitarian reasons such as hedging against disparate impact lawsuits. Nonetheless, the outputs of first-party audits rely on already existing algorithmic product development practices and software platforms. First-party audit techniques are ultimately intended to meet targets that are specified in terms of the product itself. This is why technical audits are, by design, inward-looking. Technical auditing studies how well a system performs by virtue of its own criteria for success. While those criteria *may* include protection against algorithmic harms to individuals and communities, such systems are designed to serve developers rather than the total group of people impacted by the system. In practice, this means that algorithmic impacts that can be identified and addressed inside of the development process have received the most thorough attention.

A core feature of this development process is constant iteration, with relentless tweaking of algorithmic models to find the optimal fit between training data, desired outcomes, and computational efficiency. While the model-building process is marked by metaphors of playfulness and open-endedness,[107] algorithmic governance is in tension with this playfulness, which resists formal documentation, the speed at which technology companies push out new products and services in order to remain competitive, and the need to provide accurate accounts of how systems were designed and operate when deployed. Among those involved in algorithmic governance work, it is often surprising how little technology companies actually know about the operations of their deployed models, particularly with regard to ethically relevant metadata, such as fairness parameters, demographics of the data used in training models, and considerations about geographic and cultural specificity of the training set.

And yet, many of the technical and organizational advances in algorithmic governance have come from identifying the points in the design and deployment processes that are amenable to explanation and review, and creating the necessary artifacts and internal governance mechanisms. These advances represent an emerging subset of **methods** that may need to be used by **assessors** as they conduct an AIA. As Andrew Selbst and Solon Barocas point out, the core challenge of algorithmic governance is not explaining *how* a model works, but *why* the model was designed to

106 The nascent industry of second-party algorithmic audits has already run up against some of these limits. See: Alex C. Engler, "Independent Auditors Are Struggling to Hold AI Companies Accountable," *Fast Company*, January 26, 2021, https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue; Kristian Lum and Rumman Chowdhury, "What Is an 'Algorithm'? It Depends Whom You Ask," *MIT Technology Review*, February 26, 2021, https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/.

107 Samir Passi and Steven J. Jackson, "Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects," in *Proceedings of the ACM on Human-Computer Interaction 2 (CSCW)*, 2018, 1–28, https://doi.org/10.1145/3274405.

work that way.[108] Internal audit mechanisms can therefore serve a multitude of purposes—asking *why* introduces opportunities to reflect on the proper balance between end goals, core values, and technical trade-offs. As Raji et al. have argued about internal auditing methods: "At a minimum, the internal audit process should enable critical reflections on the potential impact of a system, serving as internal education and training on ethical awareness in addition to leaving what we refer to as a 'transparency trail' of documentation at each step of the development cycle."[109]

The issue of creating a transparency trail for algorithmic systems is not a trivial problem: Machine learning models tend to shed their ethically relevant context. Each step in the technical stack (layers of software that are "stacked" to produce a model in a coordinated workflow), from datasets to deployed model, results in ever more abstraction from the context of data collection. Furthermore, as datasets and models are repurposed repeatedly, either in open repositories or between corporate departments, data scientists can be in a position of knowing relatively little about how the data has been collected and transformed as they make model development choices.[110] Thus, technical research in the algorithmic accountability field has developed documentation methods that retain ethically relevant context throughout the development process; the challenge for algorithmic impact assessment is to adapt these methods in ways that expand the scope of algorithmic harms and support the assessment of those harms as impacts.

For example, Gebru et al. (2018) proposes "datasheets for datasets," a form of documentation that could travel with datasets as they are reused and repurposed.[111] Datasheets (modeled on the obligatory safety datasheets that are included with dangerous industrial chemicals) would record the motivation, composition, context of collection, demographic details, etc., of datasets, enabling data scientists to make informed decisions about how to ethically make use of data resources. Similarly, Mitchell et al. (2019) describes a documentation process of "model cards for model reporting" that retains information about benchmarked evaluations of the model in relevant domains of use, excluded uses, and factors for evaluation, among other details.[112] Others have suggested variations of these documents specific to a domain of machine learning, such as "data statements for natural language processing," which would track the limitations

108 Andrew D. Selbst and Solon Barocas, "The Intuitive Appeal of Explainable Machines," *Fordham Law Review* 87, no. 3 (2018): 1085.

109 Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Conference on Fairness, Accountability, and Transparency* (FAT* '20), 2020, 12.

110 Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna, "Data and Its (Dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research," ArXiv Preprint, 2020, ArXiv: 2012.05345; Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," ArXiv: 2010.13561 [Cs], October 2020, http://arxiv.org/abs/2010.13561.

111 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford, "Datasheets for Datasets," ArXiv: 1803.09010 [Cs], March 2018, http://arxiv.org/abs/1803.09010.

112 Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* - FAT* '19, 2019, 220–29, https://doi.org/10.1145/3287560.3287596.

to generalizing language models to different populations.[113]

In addition to discrete documentation for datasets and models, there is also a need for describing the organizational processes required to track the complete design process. Raji et al. (2020) describe the processes needed to support algorithmic accountability throughout the lifecycle of an AI system.[114] For example, an accountability end-to-end audit might require an accounting of how and why data scientists prioritized false positive over false negative rates, considering how that decision affects downstream stakeholders and comports with the company's or industry's values standards.[115]

Ultimately, the reporting documents of such internal audits will constitute a significant bulk of any formal AIA report; indeed, it is hard to imagine a company being able to conduct a robust AIA without having in place an accountability mechanism such as that described in Raji et al. (2020). No matter how thorough and well-meaning internal accountability auditors are, such reporting mechanisms are not yet "accountable" without formal responsibility to account for the system's consequences for *those affected by it*.

SOCIOTECHNICAL EXPERTISE

While technical audits provide crucial methods for AIAs, impact assessment **methods** will need **assessors**, particularly social scientists and other critical scholars, who have long studied how understanding race, gender, and other minoritized social identities are inextricably bound with unequal and inequitable effects of sociotechnical systems.[116] This can be seen in how a groundbreaking third-party audit like "Gender Shades" brings the concept of "intersectionality" from the critical race scholarship of Kimberlé Crenshaw to bear on facial recognition technology.[117] Similarly, ethnographers and other social scientists have studied the implications of algorithmic systems for those who are made subject to them,[118] community advocates and activists have made visible the

113 Emily M. Bender and Batya Friedman,"Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics* 6 (December 2018): 587–604, https://doi.org/10.1162/tacl_a_00041.

114 Raji et al., "Closing the AI Accountability Gap."

115 Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." ArXiv:2004.07213 [Cs], April 2020, http://arxiv.org/abs/2004.07213; Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli, "Building and Auditing Fair Algorithms: A Case Study in Candidate Screening," In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event (Canada: Association for Computing Machinery, 2021), 666–77, https://doi.org/10.1145/3442188.3445928.

116 Ruha Benjamin, *Race After Technology* (New York: Polity, 2019); Browne, *Dark Matters*; Sheila Jasanoff, ed., *States of Knowledge: The Co-Production of Science and Social Order*. (New York: Routledge, 2004).

117 Kimberle Crenshaw, "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color," *Stanford Law Review* 43, no. 6 (1991): 1241, https://doi.org/10.2307/1229039.

118 Christian Sandvig, Kevin Hamilton, Karrie Karahalios, Cedric Langbort, "When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software," *International Journal of Communication* 10 (2016): 4972–4990; Zeynep Tufekci, "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency," *Colorado Technology Law Journal* 13, no. 203 (2015); John Cheney-Lippold, "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control." *Theory, Culture & Society* 28, no. 6 (2011): 164–81.

potential harms of facial recognition entry systems for residents of apartment buildings,[119] organized labor has drawn attention to how algorithmic management has reshaped the workplace, and all such work plays a crucial role in expanding the aperture of assessment practices wide enough to include as many varieties of potential algorithmic harm as possible, so they can be rendered as impacts through appropriate assessment practices. Analogously, recognition of the disproportionate environmental harms borne by minoritized communities has allowed a more thorough accounting of environmental justice harms as part of EIAs.[120]

Social science scholarship has revealed algorithmic biases that lead to new (and old) forms of discrimination. It has argued for more efforts to ensure fairness and accountability in algorithmic systems,[121] the power-laden implications of how algorithmic representations of data subjects' lives implicate them in extractive and abusive systems,[122] and explored mundane forms of sense-making and folk theories employed by data subjects in understanding how algorithms work.[123] Research in this domain has increasingly come to consider everyday experiences of living with algorithmic systems for reasons ranging from articulating agency and voice of data subjects from the bottom up[124] to formulating data-oriented notions of social justice to inform the work of data activists and assessing the impacts of algorithmic systems.[125]

While impact assessment is based on the specifications provided by organizations building these systems and the findings of external auditors who capture impacts as top-down accounts of impacts, harms need to also be assessed from the ground up. Taking the directive to design "nothing about us without us" seriously means incorporating forms of expertise attuned to lived experience by bringing

119 Moran, "Atlantic Plaza Towers Tenants Won a Halt to Facial Recognition"; Mutale Nkonde, "Automated Anti-Blackness: Facial Recognition in Brooklyn, New York," *Journal of African American Policy*, 2019–2020, 30–36.

120 Eric J. Krieg and Daniel R. Faber, "Not so Black and White: Environmental Justice and Cumulative Impact Assessments," *Environmental Impact Assessment Review* 24 no. 7–8 (2004): 667–94, https://doi.org/10.1016/j.eiar.2004.06.008.

121 See, for example, Benjamin Edelman, "Bias in Search Results: Diagnosis and Response," *Indian JL & Tech*. 7 (2011): 16–32, http://www.ijlt.in/archive/volume7/2_Edelman.pdf; Latanya Sweeney, "Discrimination in Online Ad Delivery," *Commun. ACM* 56, no. 5 (2013): 44–54, https://doi.org/10.1145/2447976.2447990; and Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016).

122 Anna Lauren Hoffmann, "Terms of Inclusion: Data, Discourse, Violence." *New Media & Society*, September 2020, 146144482095872, https://doi.org/10.1177/1461444820958725.

123 See, for example, Taina Bucher, "The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms," *Information, Communication & Society* 20, no. 1 (2017): 30–44, https://doi.org/10.1080/1369118X.2016.1154086; Sarah Pink, Shanti Sumartojo, Deborah Lupton, and Christine Heyes La Bond, "Mundane Data: The Routines, Contingencies and Accomplishments of Digital Living," *Big Data & Society* 4, no. 1 (2017): 1–12, https://doi.org/10.1177/2053951717700924; and Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin, "When Users Control the Algorithms: Values Expressed in Practices on Twitter," *Proc. ACM Hum.-Comput. Interact.* 3 (CSCW, 2019): 138:1–138:20, https://doi.org/10.1145/3359240.

124 Nick Couldry and Alison Powell, "Big Data from the Bottom Up," *Big Data & Society* 1, no. 2 (2014): 1–5, https://doi.org/10.1177/2053951714539277.

125 See, for example, Helen Kennedy, "Living with Data: Aligning Data Studies and Data Activism through a Focus on Everyday Experiences of Datafication," *Krisis: Journal for Contemporary Philosophy*, no. 1 (2018): 18–30, https://krisis.eu/living-with-data/; and Linnet Taylor, 2017. "What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally," *Big Data & Society* 4, no. 2 (2017): 1–14, https://doi.org/10.1177/2053951717736335.

communities into the assessment process and compensating them for their expertise.[126] Other forms of expertise attuned to lived experience—social science, community advocacy, and organized labor—can also contribute insights on harms that can then be rendered as measurements through new, more technical methods and metrics. This work is already happening[127] in diffused and disparate academic disciplines as well as in broader controversies over algorithmic systems, but are not yet a formal part of any algorithmic assessment or audit process. Thus, assembling and integrating expertise—from empirical social scientists, humanists, advocates, organizers, and vulnerable individuals and communities who are themselves experts about their own lives—is another crucial component for robust algorithmic accountability from the bottom up, without which it becomes impossible to assert that the full gamut of algorithmic impacts has been assessed.

126 Charlton, James I. 2004. *Nothing about Us without Us: Disability, Oppression, and Empowerment.* 3. Dr. Berkeley, CA: University of California Press.; Costanza-Chock, Sasha. 2020. "Design Justice." Cambridge, MA: MIT Press.

127 Christin, 2020. *cf.* Sloane and Moss, "AI's social sciences deficit." *Nature Machine Intelligence,* 1 no. 8 (2019): 330–331; Rumman Chowdhury and Lilly Irani, "To Really 'Disrupt,' Tech Needs to Listen to Actual Researchers," *Wired,* June 26, 2019, https://www.wired.com/story/tech-needs-to-listen-to-actual-researchers/.

COMMENSURABILITY & METHODS

Allegheny Family Screening Tool

In 2015, the Office of Children, Youth and Families (CYF) in Allegheny County, Pennsylvania published a request for proposals soliciting a predictive service to assist child welfare call screeners by assigning risk scores to reports of child abuse, which was won by a team led by social service data science experts Rhema Vaithianathan and Emily Putnam-Hornstein.[128] Typically for US child welfare services, when someone suspects that a child is being abused they call a hotline number and provide a report to child welfare staff. The call "screener" then assesses the report and either "screens in" the child, triggering an in-person investigation, or "screens out" the child based on lack of evidence or an informed judgement regarding low risk on the agency's rubric. The AFST was designed to make this decision-making process efficient. The system makes screening recommendations (but not investigative predictions nor administrative judgements) based on patterns across the linked administrative datasets about Allegheny County residents, ranging from police records, school records, and other social services.[129] Often these datasets contain information about families over multiple generations—particularly if the family is of low socio-economic status and has interacted with public services many times over decades—providing screeners with a proxy bird's-eye-view over the child's family history, and its interpretation of risk in relation to the population of similar children. Ultimately, the screening recommendation (represented as a numerical score) is a prediction answering the question of "how likely is it that a child with a statistically similar history and family background would be either the subject of a major abuse investigation or placed into foster care in the next year?"

Given the sensitivity of this data, the designers of AFST participated in a second-party algorithmic fairness audit, conducted by quantitative public policy expert Alexandra Chouldechova.[130] Chouldechova et al. is an early case study of how to conduct an audit and recalibration of an automated decision system for quantifiable demographic bias, using a "fairness aware" approach that favors predictive accuracy across groups. They further solicited two ethicists, Tim Dare and Eileen Gambrill, to conduct a second-party audit centered on the question of whether implementing AFST is likely to create the best outcomes of available alternatives, including proceeding with the status quo without any predictive service.[131] Additionally, historian Virginia Eubanks features a third-party qualitative audit of the AFST in her book *Automating Inequality*.[132]

Dare and Gambrill's ethical analysis proceeds from first principles and does not center lived experience of people interacting with the AFST as a sociotechnical system.

128 Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney, "Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation," Aukland: Centre for Social Data Analytics, Auckland University of Technology, 2017, https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf.

129 *Ibid*.

130 Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan, "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," in *Conference on Fairness, Accountability and Transparency*, 2018, 134–48, http://proceedings.mlr.press/v81/chouldechova18a.html.

131 Tim Dare and Eileen Gambrill, "Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County," in Vaithianathan et al., 2017.

132 Eubanks *Automating Inequality*..

For example, regarding the risk of algorithmic bias toward non-white families, they assume that the CYF interventions will be experienced primarily as supportive, rather than punitive: "It matters ethically … that a high risk score will trigger further investigation and positive intervention rather than merely more intervention and greater vulnerability to punitive response."[133] However, this runs contrary to Eubank's empirical qualitative findings that her research subjects experience a perverse incentive to forgo voluntary, proactive support from CYF to avoid creating another contact with the system and thus increasing their risk scores. In the course of her research, she encountered well-intended but struggling families who had a sophisticated view of the algorithmic system from the other side, and who avoided seeking some sources of assistance, in order to avoid creating records that could be used against them. Furthermore, discussing the designer's efforts to achieve predictive parity across racial groups,[134] Eubanks argues that "the activity that introduces the most racial bias into the system is the very way the model defines measurement." She locates unfairness not in a quantitative measure of predictive parity across populations, but in the epistemic circularity of machine learning applications applied to historical records of human behavior. As Eubanks points out, the predictive score is at best a proxy for likelihood of actual harm to a child—it is really a measure of how *this* community of reporters, screeners, family welfare agents, judges and juries has historically responded to children like *this*. Systemically marginal populations often find it hardest to represent themselves adequately through their data creating perverse cycles of discrimination in machine learning-based predictions.

Reading Eubanks', the ethicists', and the technologists' accounts of AFST back-to-back, one could be excused for thinking that they are describing different systems. This is not to claim that the AFST designers or CYF were unethical or sloppy. Indeed, their work is notable for exceeding the norms of technical scholarship in incorporating ethical research methods and making the ethical reasoning behind design decisions transparent. Eubanks acknowledges that CYF's approach is likely a best-case scenario for using machine learning in social services. Whatever else might be said about its consequences, the process used to create and deploy the AFST remains exemplary. This shows that the commensurability of the methods deployed in AIAs pose a significant challenge: there is no final, definitive measure of "impact." It requires a judicious cobbling together of contested evidence and conflicting perspectives under a consensus process. Assembling the right expertise and constituencies to generate legitimacy is, in the end, the only way to resolve how an AIA could be adequately concluded.

133 Dare and Gambrill, "Ethical Analysis" in Vaithianathan et al., 2017.

134 Chouldechova et al., "A Case Study of Algorithmic-Assisted Decision Making."

CONCLUSION: GOVERNING WITH AIAs

For an AIA process to really achieve accountability, a number of questions about how to structure these assessments will need to be answered. Many of these questions can be addressed by carefully considering how to tailor each of the 10 constitutive components of an impact assessment process specifically for AIAs. Like at any restaurant, a menu of options exists for each course—but it may sometimes be necessary to order "off menu." Constructing an AIA process also needs to satisfy the multiple, overlapping, and disparate needs of everyone involved with algorithmic systems.[135]

A robust AIA process will also need to lay out the scope of harms that are subject to algorithmic impact assessment. Quantifiable algorithmic harms like disparate impacts to protected classes of individuals are well studied, but there are a range of other algorithmic harms which require consideration in how impacts get assessed. These algorithmic harms include (but are not limited to) representational harms, allocational harms, and harms to dignity.[136] For an AIA process to encompass the appropriate scope of potential harms, it will need to first consider: (1) how to integrate the interests and agency of affected individuals and communities into measurement practices; and (2) the mechanisms through which community input will be balanced against the power and autonomy of private developers of algorithmic systems, and (3) the constellation of other governance and accountability mechanisms at play within a given domain.

A robust AIA process will also need to acknowledge that not all algorithmic systems may require an AIA—all computation is built on "algorithms" in a strictly technical sense, but there is a vast difference between something like a bubble-sort algorithm that is used in prosaic computational processes like alphabetizing lists, and algorithmic *systems* that are used to shape social, economic, and political life, for example, to decide who gets a job and who does not. Many algorithmic systems will not clearly fall into neat categories that either definitely require or are definitely exempt from an AIA. Furthermore, technical methods alone will not illuminate which category a system belongs in. Algorithmic impact assessment will require an accountable process for determining what **catalyzes** an AIA, based on the context and the content of an algorithmic system and its specified purpose. These characteristics may include the domain in which it operates, as above, but might also include the actor operating the system, the funding entity, the function the system serves, the type of training data involved, and so on. The proper role of government regulators in outlining requirements for when an AIA is necessary, what it consists of in particular contexts, and how it is to be evaluated, also remain to be determined.

Given the differences in impact assessment processes laid out above, and the variability of algorithmic systems and their myriad effects on the world, it is worthwhile to step back and observe how impact assessments, in general, act in the

135 Boven's definition of accountability that we have been working from throughout this report is useful, in particular because it allows us to identify five distinct forms of accountability. Knowing these distinct forms of accountability is an important step toward what forms of accountability manifest in the case of algorithmic impact assessments. They are: (a) *political accountability* for those who administer algorithmic systems in the public interest; (b) *legal accountability* for harms produced by algorithmic systems; (c) *administrative accountability* to ensure that the potential impacts of an algorithmic system are properly assessed before they are allowed to operate in the world; (d) *professional accountability* for those who build algorithmic systems to ensure that their specifications and assessments meet relevant technical standards; and finally, (e) *social accountability* through which the public can hold algorithmic systems and their operators responsible for algorithmic harms through assessment of impacts.

136 Barocas et al., "The Problem with Bias."

world. Namely, impact assessments structure power, sometimes in ways that reinforce structural inequalities and unjust hierarchies. They produce and distribute risk, they are exercises of power, and they provide a means to contest power and distribution of risk. In analyzing impact assessments as accountability mechanisms, it is crucial to see impact assessments themselves as sets of power-laden practices that instantiate and structure power at the same time as they provide a means for contesting existing power relationships. For AIAs, the ways in which various components are selected and various forms of expertise are assembled, are directly implicated in the distribution of power. Therefore, these components must be selected with an awareness of how impact assessment can at times fall short of equitably distributing power, replicating already existing hierarchies, and produce the appearance of accountability without tangibly reducing harms. With these observations in mind, we can begin to ask practical questions about how to construct an algorithmic impact assessment process.

One of the first questions that needs to be addressed is: **who should be considered as stakeholders for the purposes of an AIA?** These stakeholders could include system developers (private technology companies, civic tech organizations, and government agencies that build such systems themselves), system operators (businesses and government agencies that purchase or license systems from third-party vendors), independent critical scholars who have developed a wide range of disciplinary forms of expertise to investigate the social and environmental implications of algorithmic systems, independent auditors who can conduct thorough technical investigations into the design and behavior of algorithmic systems, community advocacy organizations that are closely connected to the individuals and communities most vulnerable to potential harms, and government agencies tasked with oversight, permitting, and/or regulation.

Another question that needs to be asked is: **What should the relationship between stakeholders be?** Multi-stakeholder actions can be coordinated through a number of means, from implicit norms to explicit legislation, and an AIA process will have to determine whether government agencies ought to be able to mandate changes in an algorithmic system developed or operated by a private company, or if third-party certification of acceptable impacts are sufficient. It will also have to determine the appropriate role of public participation and the degree of access offered to community advocates and other interested individuals. AIAs will also have to identify the role independent auditors and investigators might be required to play, and how they would be compensated.

In designing relationships between stakeholders, questions of power arise: **Who is empowered through an AIA and who is not? Relatedly, how do disparate forms of expertise get represented in an AIA process?** For example, if one stakeholder is elevated to the role of accountability forum, it is given significant power over other actors. Similarly, the ways different forms of expertise are brought into relation to each other also shapes who wields power in an AIA process. The expertise of an advocacy organization in documenting the extent of algorithmic harms is different than that of a system developer in determining, for example, the likely false positive rates of their system. Carefully selecting the components of an AIA will influence whether such forms of expertise interact adversarially or learn from each other.

These questions form the theoretical basis for addressing more practical legal, policy, and technical concerns, particularly around:

1. **The role of private industry**—those who develop AI systems for their own products and those who act as vendors to government and other private enterprises—in providing technical descriptions of the systems they build and documenting their potential, or actual, impacts.

2. **The role of independent experts** on algorithmic audit and community studies of AI systems, external auditors commissioned by AI system developers, and internal technical audits conducted by AI system developers in delineating the likely impacts of such systems.

3. **The appropriate relationship between regulatory agencies, community advocates, and private industry** in negotiating the scope of impacts to be assessed, the acceptable thresholds for those impacts, and the means by which those impacts are to be minimized or mitigated.

4. Whether **private sector and public sector** uses of algorithmic systems should be regulated by the same AIA mechanism.

5. **How to specify the scope of AIAs** to reasonably delineate what types of algorithmic systems, using which types of data, operating at what scale, and affecting which people or activities should be subject to audit and assessment, and which institutions—private organizations, government agencies, or other entities—should have authority to mandate, evaluate, and/or enforce them.

Governing algorithmic systems through AIAs will require answering these questions in ways that reflect the current configurations of resources in the development, procurement, and operation of such systems while also experimenting with ways to shift political power and agency over these systems to affected communities. These current configurations need not, and should not, be taken as fixed in stone but merely as the starting point from which the impacts to those most affected by algorithmic systems, and most vulnerable to harms, can be incorporated into structures of accountability. This will require a far better understanding of the value of algorithmic systems for people who live with them, and their evaluations of and responses to the types of algorithmic risks and harms they might experience. It will also require deep knowledge of the legal framings and governance structures that could plausibly regulate such systems, and their integration with the technical and organizational affordances of firms developing algorithmic systems.

Finally, this report points to a need to develop robust frameworks in which consensus can be developed from among the range of stakeholders necessary to assemble an algorithmic impact assessment process. Such multi-stakeholder collaborations are necessary to adequately assemble, evaluate, and document algorithmic impacts and are shaped by evolving sociocultural norms and organizational practices. Developing consensus will also require constructing new tools for evaluating impacts, and understanding and resolving the relationship between actual or potential harms and the way such harms are measured as impacts. The robustness of impacts as proxies of harms can only be maintained by bringing together the multiple disciplinary and experiential forms of expertise in engaging with algorithmic systems. After all, impact assessments are a means to organize whose voices count in governing algorithmic systems.

THE 10 CONSTITUTIVE COMPONENTS OF IMPACT ASSESSMENT

	Sources of Legitimacy	Actor(s) and Forum [2]	Catalyzing Event	Time Frame	Public Access	Public Consultation	Methods	Assessors	Impacts	Harms and Redress
Component Description	Legal or regulatory mandate	Who reports to whom?	What triggers assessment process?	Assessement conducted before or after deployment	Can public access evidence?	Is public input solicited?	Measurement practices	Who conducts assessment?	What is measured?	How are harms mitigated or minimized?
Fiscal Impact Assessments (FIA) [1]	Broad public respect for rational decision-making on the part of municipal authorities	*Actor(s):* Municipal authorities such as City Council *Forum:* Constituents who may vote out such authorities"	When a municipal government decides that it is required to evaluate a proposed project	Performed *ex ante* with usually no post hoc review	Fiscal impact reports are filed with the municipality as public record, but local regulations may vary	*Is not necessary,* but may take the form of evidence gathering through stakeholder interviews with the public	The focus is on financial accounting and assessing impacts relative to a counterfactual world in which the project does not happen	Urban Planning Office or Urban Policy Institute or Consulting firm	Assessed in terms of municipal fiscal health and sometimes the actor's ability to provide other municipal services	Potential decline in city services because of negative fiscal impact. The assessment is only intended to inform decision-making and does not account for redress
Environmental Impact Assessments (EIA)	National Environmental Protection Act of 1969 (and subsequent related legislation)	*Actor(s):* Project Developers such as an energy company *Forum:* Permitting agency such as the Environmental Protection Agency (EPA)"	When a proposed project receives federal (or certain state-level) funding or crosses state lines	Performed *ex ante* often with ongoing monitoring and mitigation of harms	Impact statements are public along with a stipulated period of public comment	*Is mandatory* with explicit requirements for stakeholder and community engagement as well as public comments	The focus is on assessing impact on the environment as a resource for communal life by assembling diverse forms of expertise and public comments	Consulting firm (occasionally a design-build firm)	Assessed in terms of changes to the ready availability and viability of environmental resources for a community	Environmental degradation, pollution, destruction of cultural heritage, etc. The assessment is oriented to mitigation and lays the groundwork for standing to seek redress in court cases
Human Rights Impact Assessments (HRIA)	The Universal Declaration of Human Rights (UDHR) adopted by the United Nations in 1948	Exhibits *actor/forum collapse* where a corporation is the actor as well as the forum [3].	When a company voluntarily commissions it or experiences reputational harm from its business practices	Performed *ex post* as a forensic investigation of existing business practices	Privately commissioned and only released to the public at the discretion of the company	*Is not necessary,* but may take the form of evidence gathering through rightsholder interviews with the public	The focus is on articulating impacts on human rights as proxies for harms already experienced through rightsholder interviews	Consulting firm	Assessed in terms of abstract conditions that determine quality of life within a jurisdiction irrespective of how harms are experienced on the ground	The impacts assessed remain distant from the harms experienced and thus, do not provide standing to seek redress. Redress remains strictly voluntary for the company.
Data Protection Impact Assessments (DPIA)	General Data Protection Regulation (GDPR) adopted by the EU in 2016 and enforced since 2018	*Actor(s):* Data controllers who store sensitive user data *Forum:* The National Data Protection Commission of any country within the EU"	When a proposed project processes data of individuals in a manner that produces high risks to their rights	Performed *ex ante,* although they are stipulated to be ongoing	Impact statements are not made public, but can be disclosed upon request	*Is mandatory* without specifying the goals the process would achieve beyond mere notification	The focus is on data management practices and anticipating impacts for individuals whose data is processed	In big companies, it is usually done internally. For smaller companies, it is conducted externally through consulting firms	Assessed in terms of how rights and freedoms of individual data subjects are impinged	Harms and redress are much more closely linked with the focus of the assessment on documenting mitigation strategies for potential harms
Privacy Impact Assessments (PIA)	Fair Information Practice Principles developed in the 1973 and codified in the Privacy Act of 1974	*Actor(s):* Any government agency deploying an algorithmic system *Forum:* No distinct forum apart from public writ large and possible fines under applicable laws"	When a proposed project or change in operation of existing systems leads to collection of personally identifiable information	Performed *ex ante,* often post-design and pre-launch with usually no post hoc review	Such assessments are public, but their technical complexity may render them difficult to understand	*Is mandatory* without specifying the goals the process would achieve beyond mere notification	The focus is on managing privacy and producing a statement on how a proposed system will handle private information in accordance with relevant law	Project Managers, Chief Privacy Officer, Chief Information Security Officer, and Chief Information Officers. Independence of assessors is mandatory.	Assessed in terms of how the actor might be impacted as a result of how individuals' privacy may be compromised by actor's data collection practices	Harms and redress are much more closely linked with the focus of the assessment on documenting mitigation strategies for potential harms

[1] This table contains general descriptions of how the components are structured within each impact assessment process. Unless specified otherwise such as in the case of DPIA, we have focussed on the jurisdictions within the United States in our analysis of impact assessment processes.

[2] In each case of impact assessments, the possibility of public censure and reputational harms because of widespread publicity of the harms of a system developed/managed by the actor remains an alternative recourse for practically achieving accountability.

[3] Corporations are made accountable on their own volition. They are often spurred to make themselves accountable because of a reputational harm they have suffered. They are not only held accountable by themselves, but also through public visibility of the accountability process. An HRIA makes public the HR impacts of a company, and sets a standard against which the company attempts to improve its impacts.

BIBLIOGRAPHY

107th US Congress, *E-Government Act of 2002*.

Ada Lovelace Institute. "Examining the Black Box: Tools for Assessing Algorithmic Systems." Ada Lovelace Institute. April 29, 2020. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.

Allyn, Bobby. "'The Computer Got It Wrong': How Facial Recognition Led To False Arrest Of Black Man." *NPR*. June 24, 2020. https://www.npr.org/2020/06/24/882683463/the-computer-got-it-wrong- how-facial-recognition-led-to-a-false-arrest-in-michigan.

Arnstein, Sherry R. "A Ladder of Citizen Participation." *Journal of the American Planning Association* 85, no. 1 (2019): 12.

Article 29 Data Protection Working Party. "Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is 'Likely to Result in a High Risk' for the Purposes of Regulation 2016/679." WP 248 rev. 1. 2017. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236.

Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. "The problem with bias: from allocative to representational harms in machine learning." *Special Interest Group for Computing, Information and Society (SIGCIS) 2017*. 2017.

BAE Urban Economics. "Connect Menlo Fiscal Impact Analysis." City of Menlo Park Website. 2016. Accessed March 22, 2021. https://www.menlopark.org/DocumentCenter/View/12112/Att-J_FIA.

Bamberger, Kenneth A., and Deirdre K. Mulligan. "PIA Requirements and Privacy Decision-Making in US Government Agencies." In *Privacy Impact Assessment*, edited by David Wright and Paul De Hert, 225–50. Dordrecht: Springer, 2012. https://link.springer.com/chapter/10.1007/978-94-007-2543-0_10.

Bartlett, Robert V. "Rationality and the Logic of the National Environmental Policy Act." *Environmental Professional* 8, no. 2 (1986): 105–11.

Bender, Emily M., and Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6 (December 2018): 587–604. https://doi.org/10.1162/tacl_a_00041.

Benjamin, Ruha. *Race After Technology*. New York: Polity, 2019.

Bock, Kristen, Christian R. Kuhne, Rainer Muhlhoff, Meto Ost, Jorg Poole, and Rainer Rehak. "Data Protection Impact Assessment for the Corona App." *Forum InformatikerInnen für Frieden und gesellschaftliche Verantwortung (FIfF) e.V.* 2020. https://www.fiff.de/dsfa-corona.

Booker, Sen. Cory. "Booker, Wyden, Clarke Introduce Bill Requiring Companies To Target Bias In Corporate Algorithms." Press Office of Sen. Cory Booker (blog). April 10, 2019. https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms.

Bovens, Mark. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal* 13, no. 4 (2007): 447–68. https://doi.org/10.1111/j.1468-0386.2007.00378.x.

Brammer, John Paul. "Trans Drivers Are Being Locked Out of Their Uber Accounts." *Them*. August 10, 2018. https://www.them.us/story/trans-drivers-locked-out-of-uber.

Browne, Simone. *Dark Matters: On the Surveillance of Blackness*. Durham, NC: Duke University Press, 2015.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." ArXiv:2004.07213 [Cs], April 2020. http://arxiv.org/abs/2004.07213.

BSR. "Human Rights Impact Assessment: Facebook in Myanmar." Technical Report. 2018. https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf.

Bucher, Taina. "The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms." *Information, Communication & Society* 20, no. 1 (2017): 30–44. https://doi.org/10.1080/1369118X.2016.1154086.

Bullard, Robert D. "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, 1999.

Buolamwini, Joy. "Amazon Is Right: Thresholds and Legislation Matter, So Does Truth." *Medium*. February 7, 2019. https://medium.com/@Joy.Buolamwini/amazon-is-right-thresholds-and- legislation-matter-so-does-truth-6cfdf6005c80.

———. "Response: Racial and Gender Bias in Amazon Rekognition—Commercial AI System for Analyzing Faces." *Medium*. April 24, 2019. https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon- rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of Machine Learning Research*. Vol. 81. 2018. http://proceedings.mlr.press/v81/buolamwini18a.html.

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

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

Bureau of Land Management. "Environmental Assessment for Anadarko E&P Onshore LLC Kinney Divide Unit Epsilon 2 POD." WY-070-14-264. Johnson County, WY: Bureau of Land Management, Buffalo Field Office. 2014. https://eplanning.blm.gov/public_projects/nepa/67845/84915/101624/KDUE2_EA.pdf.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3, no. 1 (2016). https://doi.org/10.1177/2053951715622512.

Burrell, Jenna, Zoe Kahn, Anne Jonas, and Daniel Griffin. "When Users Control the Algorithms: Values Expressed in Practices on Twitter." *Proc. ACM Hum.-Comput. Interact.* 3 (CSCW 2019): 138:1–138:20. https://doi.org/10.1145/3359240.

Cadwalladr, Carole, and Emma Graham-Harrison. "The Cambridge Analytica Files." *The Guardian*. 2018 https://www.theguardian.com/news/series/cambridge-analytica-files.

Cardoso, Tom, and Bill Curry. 2021. "National Defence Skirted Federal Rules in Using Artificial Intelligence, Privacy Commissioner Says." *The Globe and Mail*, February 7, 2021. https://www.theglobeandmail.com/canada/article-national-defence-skirted-federal-rules-in-using-artificial/.

Cashmore, Matthew, Richard Gwilliam, Richard Morgan, Dick Cobb, and Alan Bond. "The Interminable Issue of Effectiveness: Substantive Purposes, Outcomes and Research Challenges in the Advancement of Environmental Impact Assessment Theory." *Impact Assessment and Project Appraisal* 22, no. 4 (2004): 295–310. https://doi.org/10.3152/147154604781765860.

Chander, Sarah, and Ella Jakubowska. "EU's AI Law Needs Major Changes to Prevent Discrimination and Mass Surveillance." European Digital Rights (EDRi), 2021. https://edri.org/our-work/eus-ai-law-needs-major-changes-to-prevent-discrimination-and-mass-surveillance/.

Cheney-Lippold, John. "A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control." *Theory, Culture & Society* 28, no. 6 (2011): 164–81.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions." In Conference on Fairness, Accountability and Transparency. 2018, 134–48. http://proceedings.mlr.press/v81/chouldechova18a.html.

Chowdhury, Rumman, and Lilly Irani "To Really 'Disrupt,' Tech Needs to Listen to Actual Researchers." *Wired*, June 26, 2019. https://www.wired.com/story/tech-needs-to-listen-to-actual-researchers/.

Christin, Angèle. "Algorithms in Practice: Comparing Web Journalism and Criminal Justice." *Big Data & Society* 4, no. 2 (2017): 205395171771885. https://doi.org/10.1177/2053951717718855.

Cole, Luke W. "Remedies for Environmental Racism: A View from the Field." *Michigan Law Review* 90, no. 7 (June 1992): 1991. https://doi.org/10.2307/1289740.

City of New York Office of the Mayor. "Establishing an Algorithms Management and Policy Officer." Vol. EO No. 50. 2019. https://www1.nyc.gov/assets/home/downloads/pdf/executive-orders/2019/eo-50.pdf.

Clarke, Yvette D. "H.R.2231—116th Congress (2019–2020): Algorithmic Accountability Act of 2019." 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231.

Couldry, Nick, and Alison Powell. "Big Data from the Bottom Up." *Big Data & Society* 1, no. 2 (2014): 1–5. https://doi.org/10.1177/2053951714539277.

Council of Europe, and European Parliament. "Regulation on European Approach for Artificial Intelligence Laying Down a Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." 2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

Crenshaw, Kimberle. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color." *Stanford Law Review* 43, no. 6 (1991): 1241. https://doi.org/10.2307/1229039.

Dare, Tim, and Eileen Gambrill. "Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County," Alleghany County Analytics. 2017. https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf.

Dietrich, William, Christina Mendoza, and Brennan, Tim. "COMPAS Risk Scales: Demonstrating Accuracy, Equity and Predictive Parity." Northpointe Inc. Research Department. 2016. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

Edelman, Benjamin. "Bias in Search Results: Diagnosis and Response." Indian JL & Tech. 7 (2011): 16–32. http://www.ijlt.in/archive/volume7/2_Edelman.pdf.

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

Engler, Alex C. "Independent Auditors Are Struggling to Hold AI Companies Accountable." *Fast Company*. January 26, 2021. https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue.

Erickson, Jessica. "Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System." 89 *Washington Law Review* 1425 (2014): 1444–45.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press, 2018.

European Commission. "On Artificial Intelligence – A European Approach to Excellence and Trust." White Paper. Brussels. 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Federal Trade Commission. "Privacy Online: A Report to Congress." US Federal Trade Commission. 1998. https://www.ftc.gov/sites/default/files/documents/ reports/privacy-online-report-congress /priv-23a.pdf.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. "Datasheets for Datasets." ArXiv: 1803.09010 [Cs], March 2018. http://arxiv.org/abs/1803.09010.

Götzmann, Nora, Tulika Bansal, Elin Wrzoncki, Catherine Poulsen-Hansen, Jacqueline Tedaldi, and Roya Høvsgaard. "Human Rights Impact Assessment Guidance and Toolbox." Danish Institute for Human Rights. 2016.

Government of Canada. "Canada-ca/Aia-Eia-Js." JSON. Government of Canada. 2016. https://github.com/canada-ca/aia-eia-js.

Government of Canada. "Algorithmic Impact Assessment – Évaluation de l'Incidence Algorithmique." Algorithmic Impact Assessment. June 3, 2020. https://canada-ca.github.io/aia-eia-js/.

Green, Ben, and Yiling Chen. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*. 90–99. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3287560.3287563.

Hamann, Kristine, and Rachel Smith. "Facial Recognition Technology: Where Will It Take Us?" *Criminal Justice Magazine*, 2019. https://www.americanbar.org/groups/criminal_justice/publications/criminal-justice-magazine/2019/spring/facial-recognition-technology/.

Hanna. "Data Protection Advocates Prevail: Germany Builds a Covid-19 Tracing App with Decentralized Storage." *Tutanota*. April 29, 2020. https://tutanota.com/blog/posts/germany-privacy-covid-app.

Hill, Kashmir. "Wrongfully Accused by an Algorithm." *The New York Times*, June 24, 2020. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

———. "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match." *The New York Times*, December 29, 2020. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html.

Hoffmann, Anna Lauren. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society* 22 no. 7 (2019): 900–915. https://doi.org/10.1080/1369118X.2019.1573912.

———. "Terms of Inclusion: Data, Discourse, Violence." *New Media & Society*, September 2020, 146144482095872. https://doi.org/10.1177/1461444820958725.

Hogan, Libby and Michael Safi. "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.

Hutchinson, Ben, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure." ArXiv: 2010.13561 [Cs], October 2020. http://arxiv.org/abs/2010.13561.

International Association for Impact Assessment. "Best Practice." Accessed May 2020. https://iaia.org/best-practice.php.

Jasanoff, Sheila, ed. *States of Knowledge: The Co-Production of Science and Social Order*. International Library of Sociology. New York: Routledge, 2004.

Johnson, Khari. "Amsterdam and Helsinki Launch Algorithm Registries to Bring Transparency to Public Deployments of AI." *VentureBeat*, September 28, 2020. https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/.

Johnson, Scott K. "Amid Oil- and Gas-Pipeline Halts, Dakota Access Operator Ignores Court." *Ars Technica*. July 8, 2020. https://arstechnica.com/science/2020/07/keystone-xl-dakota-access-atlantic-coast-pipelines-all-hit-snags/

"JointStatement on Contact Tracing." 2020. https://main.sec.uni-hannover.de/JointStatement.pdf.

Karlin, Michael. "The Government of Canada's Algorithmic Impact Assessment: Take Two." *Medium*, August 7, 2018. https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f.

———. "Deploying AI Responsibly in Government." *Policy Options* (blog), February 6, 2018. https://policyoptions.irpp.org/magazines/february-2018/deploying-ai-responsibly-in-government/.

Kemp, Deanna and Frank Vanclay. "Human rights and impact assessment: clarifying the connections in practice." *Impact Assessment and Project Appraisal* 31, no. 2 (June 2013): 86–96. https://doi.org/10.1080/14615517.2013.782978.

Kennedy, Helen. "Living with Data: Aligning Data Studies and Data Activism through a Focus on Everyday Experiences of Datafication." *Krisis: Journal for Contemporary Philosophy*, no. 1 (2018): 18–30. https://krisis.eu/living-with-data/.

Klein, Ezra. "Mark Zuckerberg on Facebook's hardest year, and what comes next." *Vox,* April 2, 2108. https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.

Kotval, Zenia, and John Mullin. "Fiscal Impact Analysis: Methods, Cases, and Intellectual Debate." Lincoln Institute of Land Policy Working Paper. Lincoln Institute of Land Policy. 2006. https://www.lincolninst.edu/sites/default/files/pubfiles/kotval-wp06zk2.pdf.

Krieg, Eric J., and Daniel R. Faber. "Not so Black and White: Environmental Justice and Cumulative Impact Assessments." *Environmental Impact Assessment Review* 24, no. 7–8 (2004): 667–94. https://doi.org/10.1016/j.eiar.2004.06.008.

Lapowsky, Issie and Emily Birnbaum. "Democrats Have Won the Senate. Here's What It Means for Tech." *Protocol — The People, Power and Politics of Tech, January 6, 2021*. https://www.protocol.com/democrats-georgia-senate-tech.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*. Accessed March 22, 2021. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=6LHoUCqhSP02JHSsAi7mlAd73V6zJtgb.

Latonero, Mark. "Governing Artificial Intelligence: Upholding Human Rights & Dignity." Data & Society Research Institute, 2018. https://datasociety.net/library/governing-artificial-intelligence/.

———. "Can Facebook's Oversight Board Win People's Trust?" *Harvard Business Review*, January 2020. https://hbr.org/2020/01/can-facebooks-oversight-board-win-peoples-trust.

Latonero, Mark, and Aaina Agarwal. "Human Rights Impact Assessments for AI: Learning from Facebook's Failure in Myanmar." Carr Center for Human Rights Policy, Harvard Kennedy School, 2021.

Lemay, Mathieu. "Understanding Canada's Algorithmic Impact Assessment Tool." Toward Data Science (blog), June 11, 2019. https://towardsdatascience.com/understanding-canadas-algorithmic-impact-assessment-tool-cd0d3c8cafab.

Lewis, Rachel Charlene. 2020. "Making Facial Recognition Easier Might Make Stalking Easier Too." *Bitch Media*, January 31, 2020. https://www.bitchmedia.org/article/very-online/clearview-ai-facial-recognition-stalking-sexism.

Lum, Kristian, and Rumman Chowdhury. "What Is an 'Algorithm'? It Depends Whom You Ask." *MIT Technology Review*, February 26, 2021. https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. "Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 735–746. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445935.

Milgram, Anne, Alexander M. Holsinger, Marie Vannostrand, and Matthew W. Alsdorf. "Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making." *Federal Sentencing Reporter* 27, no. 4 (2015): 216–21. https://doi.org/10.1525/fsr.2015.27.4.216.

Mikians, Jakub, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. "Detecting Price and Search Discrimination on the Internet." *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI,* 79–84. Redmond, Washington: ACM Press, 2012. https://doi.org/10.1145/2390231.2390245.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19,* 220–29. 2019. https://doi.org/10.1145/3287560.3287596.

Moran, Tranae'. "Atlantic Plaza Towers Tenants Won a Halt to Facial Recognition in Their Building: Now They're Calling on a Moratorium on All Residential Use" *AI Now Institute* (blog). January 9, 2020. https://medium.com/@AINowInstitute/atlantic-plaza-towers-tenants-won-a -halt-to-facial-recognition-in-their-building-now-theyre-274289a6d8eb.

Morgan, Richard K. "Environmental impact assessment: the state of the art." *Impact Assessment and Project Appraisal* 30, no. 1 (March 2012): 5–14. https://doi.org/10.1080/14615517.2012.661557.

Morris, Peter, and Riki Therivel. *Methods of Environmental Impact Assessment*. London; New York: Spon Press, 2001. http://site.ebrarxy.com/id/5001176.

Nike, Inc. "Sustainable Innovation Is a Powerful Engine for Growth: FY14/15 Nike, Inc. Sustainable Business Report." Nike Inc., 2015. https://purpose-cms-production01.s3.amazonaws.com/wp-content/uploads/2018/05/14214951/NIKE_FY14-15_Sustainable_Business_Report.pdf.

Nissenbaum, Helen. "Accountability in a Computerized Society." *Science and Engineering Ethics* 2, no. 1 (1996): 25–42. https://doi.org/10.1007/BF02639315.

Nkonde, Mutale. "Automated Anti-Blackness: Facial Recognition in Brooklyn, New York." *Journal of African American Policy*, *Anti-Blackness in Policy Making: Learning from the Past to Create a Better Future, 2020–2021*. 2020.

Office of Privacy and Civil Liberties. "Privacy Act of 1974." *US Department of Justice*. https://www.justice.gov/opcl/privacy-act-1974.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

Panel for the Future of Science and Technology. "A Governance Framework for Algorithmic Accountability and Transparency." EU: European Parliamentary Research Service, 2019. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

Passi, Samir, and Steven J. Jackson. "Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects." *Proceedings of the ACM on Human-Computer Interaction 2 (CSCW)*: 1–28. 2018. https://doi.org/10.1145/3274405.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. "Data and Its (Dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research." ArXiv Preprint, 2020. ArXiv: 2012.05345.

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

Pink, Sarah, Shanti Sumartojo, Deborah Lupton, and Christine Heyes La Bond. "Mundane Data: The Routines, Contingencies and Accomplishments of Digital Living." *Big Data & Society* 4, no. 1(2017): 1–12. https://doi.org/10.1177/2053951717700924.

Power, Michael. *The Audit Society: Rituals of Verification*. New York: Oxford University Press, 1997.

Privacy Office of the Office Information Technology. "Privacy Impact Assessment (PIA) Guide." US Securities & Exchange Commission, 2007.

Putnam-Hornstein, Emily, and Barbara Needell. "Predictors of Child Protective Service Contact between Birth and Age Five: An Examination of California's 2002 Birth Cohort." *Children and Youth Services Review, Maltreatment of Infants and Toddlers 33, no. 8 (2011): 1337–44*. https://doi.org/10.1016/j.childyouth.2011.04.006.

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

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 12. Barcelona, ES, 2020.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." AI Now Institute, 2018. https://ainowinstitute.org/aiareport2018.pdf.

Roose, Kevin. "Forget Washington. Facebook's Problems Abroad Are Far More Disturbing." *The New York Times,* October 29, 2017. www.nytimes.com/2017/10/29/business/facebook-misinformation-abroad.html.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. "Automation, Algorithms, and Politics | When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software" *International Journal of Communication* 10 (2016): 19.

———. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." *Data and Discrimination: Converting Critical Concerns into Productive Inquiry Vol. 22*. Seattle WA, 2014.

Schmitz, Rob. "In Germany, High Hopes for New COVID-19 Contact Tracing App That Protects Privacy." *NPR*, April 2, 2020. https://www.npr.org/sections/coronavirus-live-updates/2020/04/02/825860406/in-germany-high-hopes-for-new-covid-19-contact-tracing-app-that-protects-privacy.

Seah, Josephine. "Nose to Glass: Looking In to Get Beyond." ArXiv: 2011.13153 [Cs], December, 2020. http://arxiv.org/abs/2011.13153.

Secretary's Advisory Committee on Automated Personal Data Systems. "Records, Computers, and the Rights of Citizens: Report." DHEW No. (OS) 73-94. US Department of Health, Education & Welfare. 1973. https://aspe.hhs.gov/report/records-computers-and-rights-citizens.

Selbst, Andrew D. "Disparate Impact in Big Data Policing." *SSRN Electronic Journal*, 2017. https://doi.org/10.2139/ssrn.2819182.

Selbst, Andrew D., and Solon Barocas. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87 (2018): 1085.

Shwayder, Maya. "Clearview AI Facial-Recognition App Is a Nightmare For Stalking Victims." *Digital Trends*. January 22, 2020. https://www.digitaltrends.com/news/clearview-ai-facial-recognition-domestic-violence-stalking/.

Sloane, Mona. "The Algorithmic Auditing Trap." OneZero (blog), March 17, 2021. https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.

Sloane, Mona and Moss, Emanuel. "AI's social sciences deficit." *Nature Machine Intelligence*, *1* no. 8 (2017): 330–331.

Sloane, Mona, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. "Participation Is Not a Design Fix for Machine Learning." *Proceedings of the 37th International Conference on Machine Learning*, 7. Vienna, Austria, 2020.

Snider, Mike. "Congress and Technology: Do Lawmakers Understand Google and Facebook Enough to Regulate Them?" *USA TODAY*, August 2, 2020. https://www.usatoday.com/story/tech/2020/08/02/google-facebook-and-amazon-too-technical-congress-regulate/5547091002/.

Star, Susan Leigh. "This Is Not a Boundary Object: Reflections on the Origin of a Concept." Science, Technology, & Human Values 35, no. 5 (2010): 601–17. https://doi.org/10.1177/0162243910377624.

Star, Susan Leigh, and James R. Griesemer. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social Studies of Science 19, no. 3 (1989): 387–420. https://doi.org/10.1177/030631289019003001.

Stevenson, Alexandra. "Facebook Admits It Was Used to Incite Violence in Myanmar." *The New York Times*, November 6, 2018. https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

Sweeney, Latanya. "Discrimination in Online Ad Delivery." *Commun. ACM* 56, no. 5 (2013): 44–54. https://doi.org/10.1145/2447976.2447990.

Tabuchi, Hiroko, and Brad Plumer. "Is This the End of New Pipelines?" *The New York Times*, July, 2020. https://www.nytimes.com/2020/07/08/climate/dakota-access-keystone-atlantic-pipelines.html.

Taylor, Linnet. "What Is Data Justice? The Case for Connecting Digital Rights and Freedoms Globally." *Big Data & Society* 4, no. 2 (2017): 1–14. https://doi.org/10.1177/2053951717736335.

Taylor, Serge. *Making Bureaucracies Think: The Environmental Impact Statement Strategy of Administrative Reform.* Stanford, CA: Stanford University Press, 1984.

Thamkittikasem, Jeff. "Implementing Executive Order 50 (2019), Summary of Agency Compliance Reporting." City of New York Office of the Mayor, Algorithms Management and Policy Officer, 2020. https://www1.nyc.gov/assets/ampo/downloads/pdf/AMPO-CY-2020-Agency-Compliance-Reporting.pdf.

"The Radical AI Podcast." *The Radical AI Podcast*, June 2020. https://www.radicalai.org/e15-deb-raji.

Treasury Board of Canada Secretariat. "Directive on Automated Decision-Making." 2019. https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

Tufekci, Zeynep. "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency." *Colorado Technology Law Journal* 13, no. 203 (2015).

United Nations Human Rights Office of the High Commissioner. "Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework." New York and Geneva: United Nations, 2011. https://www.ohchr.org/Documents/Publications/ GuidingPrinciplesBusinessHR_EN.pdf.

Wagner, Ben. "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?" *Being Profiled*, edited by Emre Bayamlioglu, Irina Baralicu, Liisa Janseens, and Mireille Hildebrant. 84–89. *Cogitas Ergo Sum: 10 Years of Profiling the European Citizen.* Amsterdam University Press, 2018. https://doi.org/10.2307/j.ctvhrd092.18.

Wieringa, Maranke. "What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372833.

Wilson, Christo, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. "Building and Auditing Fair Algorithms: A Case Study in Candidate Screening." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–77. Virtual Event, Canada: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445928.

World Food Program. "Rohingya Crisis: A Firsthand Look Into the World's Largest Refugee Camp." *World Food Program USA* (blog), 2020. Accessed March 22, 2021. https://www.wfpusa.org/articles/rohingya-crisis-a-firsthand-look-into-the-worlds-largest-refugee-camp/.

Wright, David and Paul De Hert. "Introduction to Privacy Impact Assessment." *Privacy Impact Assessment*, edited by David Wright and Paul De Hert. 3–32. Dordrecht: Springer, 2012. https://link.springer.com/chapter/10.1007/978-94-007-2543-0_1.

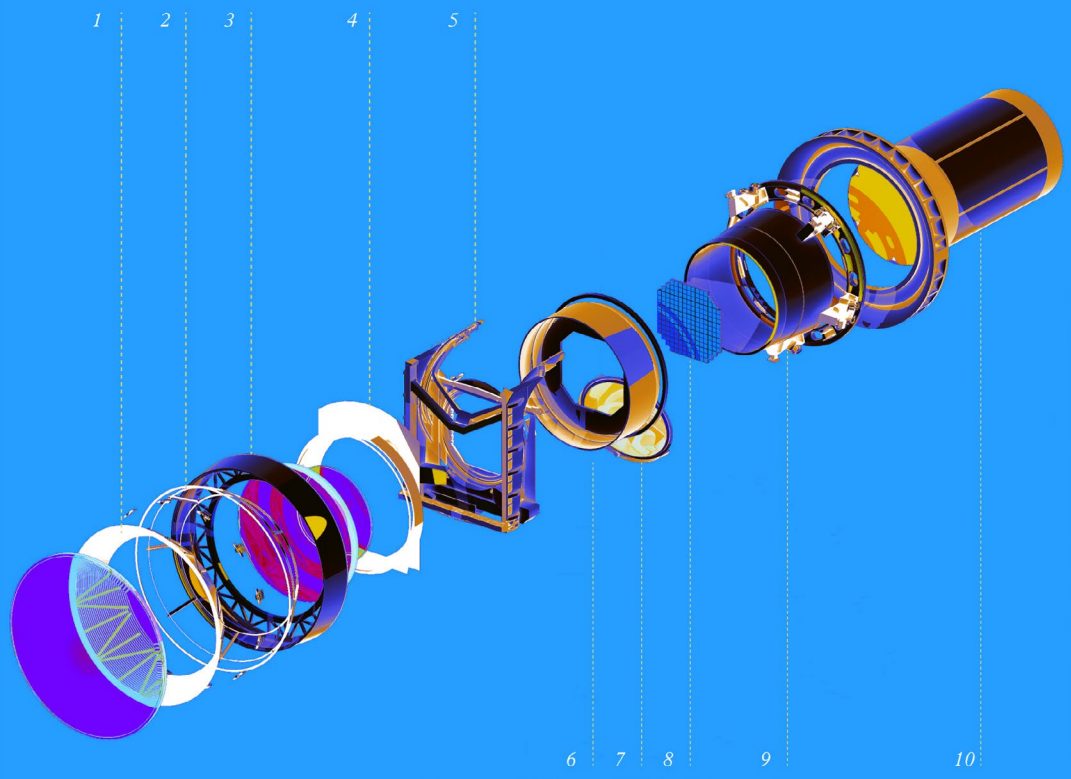
Vaithianathan, Rhema, Tim Maloney, Emily Putnam-Hornstein, and Nan Jiang. "Children in the Public Benefit System at Risk of Maltreatment: Identification via Predictive Modeling." *American Journal of Preventive Medicine* 45, no. 3 (2013): 354–59. https://doi.org/10.1016/j.amepre.2013.04.022.

Vaithianathan, Rhema, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. "Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation." Aukland: Centre for Social Data Analytics, Auckland University of Technology, 2017. https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf.

ACKNOWLEDGMENTS

This project took a long and winding path, and many people contributed to it along the way. First, we would like to acknowledge Andrew Selbst, who helped launch this project prior to moving on to a university position, and whose earlier work initialized this conversation in the scholarship. We would also like to thank Mark Latonero, whose early input was integral to developing the research presented in this report. We are especially grateful to our external reviewers, Andrew Strait and Mihir Kshirsagar, for their helpful guidance. We are also grateful to anonymous reviewers who read portions of the research in academic venues. As always, we would like to thank Sareeta Amrute who read through multiple drafts and always found the through-line to focus on. Data & Society's entire production, policy, and communications crews produced valuable input to the vision of this project, especially Patrick Davison, Chris Redwood, Yichi Liu, Natalie Kerby, Brittany Smith, and Sam Hinds. We would also like to thank The Raw Materials Seminar at Data & Society for reading much of this work in draft form. Additionally, we would like to thank the REALML community and their funder, MacArthur Foundation, for hosting important and generative conversations early in the work. We would additionally like to thank the Princeton Center for Information Technology Policy for supporting the contributions of Elizabeth Anne Watkins to this effort.

This work was funded through the Luminate Foundation's generous support of the AI on the Ground Initiative at Data & Society. This material is based upon work supported by the National Science Foundation under Award No.1704425, through the PERVADE Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Data & Society is an independent nonprofit research institute that advances new frames for understanding the implications of data-centric and automated technology. We conduct research and build the field of actors to ensure that knowledge guides debate, decision-making, and technical choices.

www.datasociety.net
@datasociety

Designed by Yichi Liu
June 2021