Patterns

CellPress
OPEN ACCESS

Perspective

Obligations to assess: Recent trends in AI accountability regulations

Serena Oduro,[1] Emanuel Moss,[2] and Jacob Metcalf[1,*]
[1]Data & Society Research Institute, New York, NY 10011, USA
[2]Intel Labs, Hillsboro, OR 97124, USA
*Correspondence: jake.metcalf@datasociety.net
https://doi.org/10.1016/j.patter.2022.100608

THE BIGGER PICTURE Recently proposed legislation would reshape how developers build algorithmic systems, requiring more documentation that includes information concerning potential bias and discriminatory impact and other types of potential harms. While such documentation regimes go by a variety of names, what they have in common is an obligation to assess the consequences of algorithmic systems beyond the technical parameters that are most familiar to engineers and system developers.

To better comply with forthcoming regulations, new techniques for documenting algorithmic system development processes are needed, as is consensus about methods for doing compliance work that centers the public interest. Practical, scalable methods that incorporate multiple forms of expertise are sorely needed to address the concerns raised in this perspective. Additionally, any move towards adopting more robust assessment and transparency practices will enable an entire ecosystem of accountability relationships to emerge.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Policymakers are increasingly turning toward assessments of social, economic, and ethical impacts as a governance model for automated decision systems in sensitive or regulated domains. In both the United States and the European Union, recently proposed legislation would require developers to assess the impacts of their systems for individuals, communities, and society, a notable step beyond the technical assessments that are familiar to the industry. This paper analyzes four examples of such legislation in order to illustrate how AI regulations are moving toward using accountability documentation to address common AI accountability concerns: identifying and documenting harms, public transparency, and anti-discrimination rules. We then offer some insights into how designers of automated decisions systems might prepare for and respond to such rules.

INTRODUCTION

Policymakers, community advocates, and some technology companies have recently turned toward proposed processes to assess the impacts of algorithmic systems as a way to introduce an additional accountability mechanism for artificial intelligence (AI) and machine learning (ML) technologies.[1–3] Although efforts to produce algorithmic accountability have taken many forms—algorithmic impact assessments (AIAs), conformity assessments, fairness audits, etc.—and make use of a variety of regulatory structures and voluntary commitments, they share some common features that are increasingly important for developers of automated decision systems and machine learning platforms to understand and address. This perspective intends to inform developers about how these trends in algorithmic accountability policy may impact AI development by describing and critically assessing a handful of recently proposed and enacted regulations.

Despite algorithmic accountability tools being in their infancy, policymakers have called for the use of impact assessments and other transparency mechanisms within proposed algorithmic accountability regulation. Impact assessments are one such emerging mechanism within the algorithmic accountability policy sphere, the aim of which is to identify potential harms or discrimination before or during an algorithmic system's development and deployment.[4] Impact assessments have been used in other policy arenas, such as urban planning, environmental policy, human rights, data protection, and government agencies and companies that manage private information, to produce accountability in varying degrees and forms.[1] Impact assessments and similar accountability instruments are particularly useful at making clear what the expected trade-offs are when

a decision may have unanticipated outcomes, when a system is substantially complex and poses many possible trade-offs, and when there are many stakeholders with potentially competing interests. The function of these accountability instruments is foremost to provide a common basis for deliberating about the outcomes of a decision and the relative desirability of those outcomes. Impact assessment processes are always co-constituted with accountability regimes: the measurement of "impact" implies an established accountability relationship in which at least one party is responsible to prevent and repair harms.[5] Facilitating contestation over the scientific facts and measurement methods—and ultimately creating a somewhat stable consensus about appropriate measurement practices—is a central feature of impact assessment.[6]

The analysis that follows illustrates a trend within algorithmic accountability policy that uses impact assessments and similar documentation as a governance mechanism. If this trend continues, those who develop and deploy algorithmic systems will need to also invest in the capacity to measure and track the consequences of their systems in the world, outside of the technical parameters of the systems themselves. This trend has implications for both the technical systems developers use to build and understand machine learning models and the organizational practices at companies that build and sell these systems. Computer scientists and engineers are well practiced in auditing the functioning of technical systems according to parameters from these disciplines: error rates, interoperability, efficiency, how humans interact with the system, etc. However, emerging policy approaches expect developers to routinely measure the functioning of their systems several steps beyond the parameters of the system. As we discuss later, while expanding the scope of assessment has the potential to make algorithmic systems more just, this also risks creating a perverse loop in which access to justice for an unrecognized group first depends on recognition as a group.[7] The expectation to consider longer range impacts on individuals, communities, and society has appeared in proposed and existing regulations and legislation, including in the United States and the European Union (EU).

As algorithmic systems make ever greater incursions into core social, political and economic aspects of our lives, their effects on our lives demand greater scrutiny. If mathematical equations will be contributing to criminal justice proceedings, then developers should be able to meet existing expectations about the transparency of evidence in courts.[8] If access to financial credit heavily depends on automated predictions of one's propensity to repay loans, then policymakers' efforts to seriously consider whether and how to hold developers accountable are called for. Justifying whether and how these kinds of systems make decisions, and whether those decisions are fair, is essential for meeting the public's expectations for accountable, safe, and trustworthy technologies. Accordingly, developers should be required to justify how those systems meet public expectations of fair decision-making processes.[9] These expectations are rarely met, in part because technical systems are often not architected to routinely provide the necessary data to answer these expectations, and the organizational structures of technology companies do not yet accommodate and incentivize such work.

RECENTLY ENACTED AND PROPOSED REGULATIONS

In this paper, we chose to highlight four recently proposed and/or enacted bills and regulations. Although this list is far from exhaustive of recent proposals to regulate the algorithmic technology sector, these four bills illustrate a wide range of algorithmic accountability legislation approaches that impact developers and represent regulatory efforts at different jurisdictional levels (regional, national, state/provincial, and local). In the following discussion, we examine these bills in terms of their potential consequences for three major themes of algorithmic accountability: identifying and documenting possible harms, public transparency, and anti-discrimination and disparate impact.

These bills would impose new requirements on developers to explore and justify consequences of their systems outside of their technical parameters:

- The Algorithmic Accountability Act of 2022 (AAA), which was introduced in February of 2022 and as of this paper has been referred to the Committee on Commerce, Science, and Transportation, is an update to one of the first major AI regulatory bills presented in the U.S. Congress.[10] Compared with its 2019 version and other legislation, the 2022 version has more robust impact assessment requirements that will challenge developers and regulators.
- New York City's Int. 1894, which was passed in New York City's City Council at the end of 2021, requires bias audits of automated employment decision systems.[11]
- California's Assembly Bill 13 (AB 13), which has been held under submission in the California Senate since August 2021, highlights the important route that procurement processes provide for algorithmic accountability and impacts developers by recommending impact assessments during state agencies' procurement process.[12]
- Finally, the European Union AI Act (EU AI Act), which was proposed in April 2021, would impact transparency documentation processes across the world as companies and entities abide by the EU AI Act's risk-framework requirements grounded in human rights impacts.[13]

These bills, which have not yet all been passed or entered into law, provide examples of current regulatory thinking to inform developers on how varying movements in transparency documentation requirements within American and European algorithmic accountability policy will not only impact their work but the importance of developer input into the creation of algorithmic accountability policy to ensure that legislation is actionable and beneficial for historically marginalized communities.

Between drafting this article and publication, the U.S. Congress has also begun considering the 2022 American Data Privacy and Protection Act (ADPPA), which directly incorporates many of the algorithmic impact assessment components of the AAA, nested inside a general data protection law. News reporting has indicated that the ADPPA has far wider support in Congress than the AAA. Given the similarity of the proposed obligations, we anticipate that the consequences of both bills would be similar from the perspective of developers concerned with algorithmic assessments.[14]

Algorithmic Accountability Act of 2022 Impact Assessment Requirements

- Evaluation of past augmented critical decision processes in the case of a new one being created.
  - Including what changes are being made, are there any past known harms, and the intended purposes of the augmented critical decision process
- Document consultation with relevant stakeholders, including whether any recommendations from stakeholders were made and whether and why the recommendations were incorporated or ignored
- Perform ongoing privacy testing
- Perform ongoing testing relating to the performance of the automated decision system or augmented critical decision process
  - Including "differential performance depending on consumers' race, sex, gender, age, disability, religion, socioeconomic or veteran status, or any other characteristics the Commission deems appropriate, including information about how such characteristics were identified in the data (including through the use of proxy data, such as zip codes)[1]
- Perform an ongoing review of industry best practices and proposals and publications from a range of experts on documented harm or better methods to develop or perform impact assessments
- Assess the need for guardrails for the automated decision system or augmented critical decision process
- Provide ongoing documentation on data or other inputs used to develop, test, maintain, or update the ADS or ACDP
- Examine the rights of consumers, including the transparency, explainability, contestability, and opportunity for recourse
- Identify likely negative impacts on consumers and mitigation strategies
  - Including which negative impacts were left unmitigated and explaining the non-discriminatory interest and why there is not another path to satisfy those means
- Describe any ongoing documentation of the development and deployment process
- Identify and describe potential stakeholder engagement processes beneficial to the development of the ADS, ACDP, or the impact assessment
- Describe which Impact assessment requirements were fulfilled and which were not conducted, including why if they were not conducted.

Figure 1. Algorithmic Accountability Act of 2022 impact assessment requirements

IDENTIFYING AND DOCUMENTING ALGORITHMIC HARMS

Algorithmic impact assessments offer the opportunity to document potential harms and discrimination before the design, acquisition, and deployment of algorithmic systems. Whereas technical research in algorithmic accountability has developed methods that document ethically relevant context through the development process, it still remains an open challenge to adapt or buttress these methods in ways that attend to the sociomaterial harms of algorithmic systems and to do so in a manner that facilitates public contestation over the trade-offs inherent in these systems.

The Automated Decision Systems Accountability Act of 2021 (AB 13), introduced in California in 2021, encouraged a bid response from any prospective contractor during the procurement process for automated decision systems for California public state agencies to submit an automated decision system impact assessment report. The disclosures required by the report center around identifying the algorithmic system, assessing its capabilities and limitations (including the system's scope of use), explaining how the system functions (including the relationship, previous risk assessment attempts and mitigation strategies), reporting previous testing and plans for future testing identifying potential disparate impacts on protected characteristics, and providing best practices for the use of the system to avoid or minimize disparate impact based on protected classes. (Under California's Unruh Civil Rights Act, protected classes include sex, race, color, religion, ancestry, national origin, disability, medical condition, genetic information, marital status, sexual orientation, citizenship, primary language, and immigration status.)[15]

The U.S. federal government has also begun to turn toward impact assessments as a way to facilitate the documentation of algorithmic harms in advance of deployment, thereby creating an incentive to revise the systems. The updated Algorithmic Accountability Act of 2022 would require impact assessments when companies are using automated systems to make critical decisions, providing consumers and regulators with much needed clarity and structure around when and how these kinds of systems are being used. This would include when companies use or develop automated decision systems or augmented critical decision processes that affect consumers' lives relating to education, employment, essential utilities, family planning, financial services, health care, housing or lodging, legal services, and any comparable legal or significant categories decided through rulemaking. Mandatory reporting and structured guidelines have made this version stronger than its 2019 iteration.

The impact assessment requirements within the Algorithmic Accountability Act of 2022 include a wide variety of assessment tasks that would require additional attention from developers (see Figure 1).

The AAA's requirement to consult relevant stakeholders (unless prohibitively difficult) provides an opportunity to recognize and address potential discrimination outside of a statistical frame.[10] The requirement for covered entities to consult relevant stakeholders, includes "internal stakeholders (such as employees, ethics teams, and responsible technology teams) and independent external stakeholders, including auditors, representatives of and advocates for impacted groups, civil society and advocates, and technology experts, as frequently as necessary."[10] The inclusion of representatives of and advocates for impacted groups provides an important opportunity for affected communities to participate in how discrimination and harms are conceived of, constructed, and addressed, albeit as refracted through advocacy organizations' concerns. By including

language that encourages this kind of consultation, the bill situates the role of developers alongside a wider range of stakeholders. This expands the possibilities for effectively identifying and preventing discrimination.

In addition, since covered entities control the impact assessment process in the AAA, it will be important for their internal teams to have socio-technical and public engagement experts who are able to support the process of identifying and consulting around potentially discriminatory impacts. Not only should protected classes be consulted but those at the intersection of these and more identities.

The EU AI Act differs from the AAA in some important ways, most notably the explicit grounding in core human rights commitments of the EU, the outright prohibition of certain systems deemed to excessively threaten those rights, and the use of a triage system that sorts proposed systems into different categories of oversight. And, as critics have noted, the EU AI Act also has much less focus on assessing the interests of impacted communities.[16] However, from the perspective of a developer, there are core similarities between the AAA's ''impact assessments'' and the EU AI Act's ''conformity assessments'' in terms of what features of a system must be tracked and assessed during development and after deployment. A recently released conformity assessment tool called capAI, released by researchers from the Oxford Internet Institute in anticipation of passage of the EU AI Act, has many of the same basic parameters as existing algorithmic impact assessment tools and pilot studies, illustrating the overlap in these approaches.[17]

Identifying harms in a manner that prevents discrimination for historically marginalized communities will require coordination among technologists, companies, legal experts, policymakers, socio-technical experts, and historically marginalized communities. Challenges to foregrounding the public interest can be addressed through multi-stakeholder convenings and community involvement in the development of transparency documentation requirements, as long as this involvement is pursued equitably and reflexively.

PUBLIC TRANSPARENCY

Commonly circulating theories of governance posit that accountability can be achieved through greater transparency. Taking the adage that ''sunlight is the best disinfectant'' to heart, many regulatory approaches require certain documents, or certain aspects of decision making, be made available to the public so that they can render judgment through democratic means. The trend of algorithmic accountability policies requiring that the results of impact assessments and audits be made public, or that users should be able to request and receive information from companies about algorithmic decisions pertaining to themselves, is in line with this dimension of public accountability. However, we note that operationalizing transparency requirements and impact assessments can often alter the organizational processes and business decisions of developers in ways that have wider reaching effects.[3]

NYC's Int. 1894 requires that all applicants or employees screened by an automated employment decision tool must receive notifications from the employer or employment agency about:

1. If an automated employment decision tool will be used during their hiring or screening process
2. The job qualifications and characteristics used in the assessment process by the automated employment decision tool
3. The opportunity for the candidate or employee to request information about the data collected, source of the data, and the employer or employment agency's data retention policy if it is not disclosed on the employer or employment agency's website

To satisfy these requirements, the employer or employment agency must keep track of what algorithms and machine learning models it is using, how many it is using, and whether the algorithm(s) it is using relies on other algorithms. Developers must also ensure that the variables the algorithms use are explainable, which can become more difficult depending on a models' type and size.[18] In addition, what information about the data is deemed important for an employee or candidate to receive and which important factors can be reported is contested. Although approaches are being developed to answer these questions (see, e.g., Data Nutrition Project) these requirements have not been standardized in the U.S. policy or legal realm.

Despite the AAA's clearly outlined requirements for summary reports (a subset of the information required in the impact assessment reports made available for the public) that will be made available in a public repository by the Federal Trade Commission (FTC), how best to select the information contained in these reports and how to report on it in an understandable manner will be a problem that requires a multi-stakeholder approach. However, compared with the 2019 version of the AAA, which made the publication of impact assessments optional, the updated version's requirement to publish summary reports of the impact assessments is another example of the growing emphasis on public transparency as a type of accountability. The AAA also requires annual reports with trends, lessons, and statistics about information gained from the impact assessments to be published by the FTC, an opportunity to foster transparency across the industry and provide regulators and researchers a broad look into how these systems are being designed and vended.

ANTI-DISCRIMINATION AND DISPARATE IMPACT MEASUREMENTS

After the rise of interest in algorithmic fairness because of well-known cases of algorithmic discrimination, the field of algorithmic fairness has grown exponentially.[19] Indeed, much of the literature on algorithmic fairness considers statistical techniques for measuring and ameliorating differential performance of algorithmic systems on different demographic groups. Likewise, transparent reporting of statistical measures of algorithmic fairness are a key component of all of the accountability mechanisms discussed here.

In the U.S. legal system, the predominant theory for understanding the type of unintentional discriminatory outcomes common to algorithmic systems is known as ''disparate impact.'' Disparate impact is a jurisprudential and statistical concept

that is intended to limit the gap between either positive or negative outcomes on the basis of protected characteristics (e.g., race/ethnicity, gender, religion), even if the policy or practice under discussion is not intentionally discriminatory. The most common formulation is the "four-fifths rule," which states that a selection rate of less than four-fifths (80%) of minority/historically disadvantaged populations in comparison with majority/historically advantaged populations is by default considered suspect and likely requires adjustment.

The four-fifths rule (and structurally similar variations) has become the dominant model for understanding discrimination in the algorithmic fairness literature, open-source fairness tools, and internal practices at technology companies. Despite the lack of a clear empirical grounding,[20] this model presents discriminatory practices in a manner that is amenable to machine learning tools and relatively easy to solve using statistical techniques. For the most part, the proposed legislation under discussion here adopts disparate impact as a central assessment practice to be included in transparency reporting.

In AB 13, disparate impact is the central anti-discrimination legal concept that would define and guide the definition and approach to measuring discrimination. AB 13 requires that contractors describe potential disparate impacts for protected classes, describe their internal policies to identify potential disparate impacts from the proposed use of the system, and provide best practices to minimize and avoid disparate impacts. Disparate impact occurs when an actor employs facially neutral policies or practices that have an adverse effect or impact on a member of a protected class.[21] However, as current anti-discrimination law has not addressed the unique and complex ways in which algorithmic discrimination manifests there are no guidelines for developers on how to analyze the discriminatory effects of their algorithms nor if it violates current anti-discrimination norms.

Disparate impact norms will also influence the implementation of the AAA in a manner that developers must be aware of. Although the AAA does not directly use the term "disparate impact," disparate impact language is interwoven throughout the bill. For example, as noted in Figure 1, the AAA calls for developers to perform testing related to differential performance between protected classes and any other characteristics the commission adds. Whether the method to test for differential performance is the four-fifths rule or another statistical method of fairness is unclear. However, the lack of clarity within the bill is not a failure but signals a future endeavor that developers, lawyers, policymakers, socio-technical experts, and impacted communities must engage with to ensure that discrimination is defined and measured in an accurate, robust, and intersectional manner.

The EU AI Act approaches discrimination and disparate impact in reference to fundamental rights individuals hold to be protected from discrimination. Consequently, assessment practices mandated by this act, specifically the required conformity assessments that document compliance with the act, are tasked with ensuring that AI systems do not discriminate on the basis of gender, race, or other demographic traits. However, many European data collection practices explicitly bar the collection of data about demographic traits that would enable robust analyses of discriminatory effects of AI systems. This presents a future challenge for both developers and regulators who wish to demonstrate compliance with non-discrimination requirements.[13] A European Commission working group has provided some preliminary recommendations on acceptable methods to measure discrimination, particularly when ethnic classificatory data are not already held by the company, noting the high-level principle that "no data collection activity should create or reinforce existing discrimination, bias or stereotypes and that the data collected should be used for the benefit of the groups they describe and society as a whole."[22]

If algorithmic impact assessments are constructed mainly within the narrow frame of technical and legal analysis of discrimination, the harms to historically marginalized communities will most likely not be identified to the fullest extent possible. With few exceptions,[23–25] the field of algorithmic fairness has focused on statistical approaches to fairness that limits the scope of analyzing harms to the technological components of an AI system instead of also addressing the societal constructs that inform the development and purpose algorithmic systems serve.[21] Focusing on statistical approaches to fairness without accounting for the fundamentally social and political nature of "race" and other protected attributes from a critical, sociological perspective can reinforce systemic oppression. For example, many methodologies within statistical approaches to algorithmic fairness conceive of race as a fixed attribute instead of as socially constructed.[26,27] The conflation of race as a fixed attribute instead of as a social construct wrongfully validates the idea that the absence or deletion of race in data collection or idea creation means that the system cannot be racist. Even more practically speaking, measuring race as a fixed construct will lead to unforeseen problems when trying to measure system performance in different global jurisdictions and cultures that may conceive of protected traits very differently. For example, the United States, Europe, and Brazil all operate with divergent and possibly incompatible understandings of race that may confound attempts to use a single measurement of system performance for the many platforms that operate in all three regions.

IMPLICATIONS FOR PRACTITIONERS

The shift toward assessment as a governance mechanism is poised to affect algorithmic development practices inside the technology industry. Developers are increasingly eager to anticipate these regulatory approaches, and broad-based popular calls for more transparency have recently led to increased interest in adopting impact assessment techniques internally.[28] These regulations will ask developers to shift their practices in significant ways. Some of these shifts may seamlessly integrate with existing practices,[29] whereas others will be more drastic departures from the status quo. Developers are already interested in producing more explainable models, or adding explainability and interpretability layers on top of otherwise inscrutable algorithms, for practical reasons. It is therefore not an extreme departure for them to pursue more explainable algorithms in order to comply with regulations like New York City's Int. 1894.

Based on our prior research into impact assessment frameworks in adjacent industries,[1] industry practitioners will need to develop new organizational and technical practices as well. Some of these new practices will merely extend longer standing activities; e.g., the EU AI Act will require developers to extend

compliance practices by contributing to ''conformity assessments'' with human-rights-based analysis that may require novel documentation but are on a continuum with similar regulations already in play. But altogether novel practices will be needed for understanding the impacts of algorithmic systems to nonusers, those who do not directly use a product but are nevertheless affected by it. New practices will also be needed for documenting impact-relevant specifications of proprietary systems, which often cannot be fully opened to public scrutiny, with a public that has a legitimate interest in the impacts produced by such systems. Finally, new organizational and institutional relationships will be needed between developers of algorithmic systems, professionals who can undertake audit and impact assessment work to compile the necessary documentation for compliance with the regulations discussed above, and communities that are most likely to be affected by algorithmic systems.

Although we have focused on formal regulatory efforts here, we also note emerging varieties of ''soft power'' that may be useful for developers seeking to understand how to conduct such assessments. The U.S. National Institute of Standards and Technology (NIST) recently released a special publication providing guidance about detecting and measuring AI bias; such reports are often prelude to a more robust standard.[30] In the face of difficulties passing legislation, this soft power is a necessary supplement. Similarly, the Institute of Electrical and Electronics Engineers (IEEE) has sponsored multiple standards regarding ethical applications of data technologies, which developers can use to align concepts and measurement practices in assessments.[31] Research institutes have also pilot-tested some AIA toolkits[16] and conformity assessment tools based on AI harms reporting.[17] Finally, several government agencies have introduced AIA tools, including Canada's Treasury Board[32] and the U.S. Office of the Chief Information Officer.[33] Although these tools are fairly basic—and not enforced by robust regulatory powers—they do point toward the types of questions obligations to assess algorithmic systems may ask.

CONCLUSION

There is a trend toward impact assessment as a governance mechanism in the regulation of algorithmic systems. Although not all of these bills will become laws, and the details of how they would be enforced are as yet unclear, this is an area that requires careful attention and consideration from AI developers. Such assessments generate shared ground truths about the functioning of a system and what consequences it may have in the world. This would enable other forms of accountability to be enacted, such as enforcement of existing norms and rules, contestation of the desirability of social and economic outcomes, and challenges over what counts as adequate due diligence and protection of vulnerable communities. Developers and deployers of automated decision systems—particularly in sensitive and regulated domains, such as medicine, education, law enforcement, criminal justice, financial services, etc.—should prepare for the technical needs and organizational practices necessary to carry out such assessments. Similarly, researchers from both the technical and social sciences should begin building consensus about what methods for measuring the impacts of such systems are needed in order to foreground the interests of impacted people and communities.

ACKNOWLEDGMENTS

The authors wish to thank Brittany Smith and Jenna Burrell, in addition to anonymous peer reviewers, for their helpful comments on this article. This research was supported in part by NSF award 1704425.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Moss, E., Watkins, E.A., Singh, R., Elish, M.C., and Metcalf, J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest (Data & Society Research Institute). http://datasociety.net/library/assembling-accountability/.

2. Ada Lovelace Institute (2022). Algorithmic Impact Assessment: A Case Study in Healthcare (Ada Lovelace Institute). https://www.adalovelaceinstitute.org/report/algorithmic-impactasssessment- case-study-healthcare.

3. Selbst, A.D. (2021). An institutional View of algorithmic impact assessments. Harv. J. Law Technol. *35*, 117–191.

4. Ada Lovelace Institute; Open Government Institute (2021). Algorithmic Accountability for the Public Sector, p. 70. https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

5. Metcalf, J., Moss, E., Watkins, E.A., Singh, R., and Elish, M.C. (2021). Algorithmic impact assessments and accountability: The co-construction of impacts. In Proceedings of the ACM Conference on Fairness, Accountability and Transparency (ACM). Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3736261.

6. Richard, A.M. (2021). Countering documents with documents'': The politics of independent environmental auditing in Mexico. PoLAR *44*, 223–239. https://onlinelibrary.wiley.com/doi/10.1111/plar.12445.

7. Tekin, Ş. (2014). The missing self in hacking's looping effects. In Classifying Psychopathology: Mental Kinds and Natural Kinds, H. Kincaid and J.A. Sullivan, eds. (The MIT Press). Philosophical psychopathology.

8. (2022). Brief of Amicus Curiae Electronic Privacy Information Center (EPIC) in Support of Apellant, Rodriguez v. Massachusetts Parole Board (Commonweath of Massachusettes Supreme Judicial Court). https://epic.org/documents/rodriguez-v-massachusetts-parole-board/.

9. Martinez, E., and Kirchner, L. (2021). The Secret Bias Hidden in Mortgage-Approval Algorithms (The Markup). https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms.

10. Wyden, B. (2021). Algorithmic Accountability Act of 2021 (S.L.C), p. 47.

11. Cumbo, L.A. (2020). The New York City Council - File #: Int 1894-2020, pp. 1894–2020. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search.

12. Chau, E. (2020). Assembly Bill 13. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB13.

13. European Commission (2019). Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011 (Text with EEA relevance.). http://data.europa.eu/eli/reg/2019/1020/oj/eng.

14. Gaffney, J.M., Holmes, E.N., and Linebaugh, C.D. (2022). Overview of the American Data Privacy and Protection Act, H.R. 8152 (Congressional Research Service), p. 5. https://crsreports.congress.gov/product/pdf/LSB/LSB10776.

15. State of California (2016). Civil Code Section 51. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=51.

16. Ada Lovelace Institute (2022). Policy Briefing: 18 Recommendations to Strengthen the EUAI Act (Ada Lovelace Institute). https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Policy-briefing-18-recommendations-to-strengthen-the-EU-AI-Act-final.pdf.

17. Floridi, L., Holweg, M., Taddeo, M., Silva, J.A., Mökander, J., and Yuni, W. (2022). capAI: A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligene Act.

18. Strandburg, K.J. (2019). Rulemaking and inscrutable automated decision tools. Columbia Law Rev. *119*, 1851–1886.

19. Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., Manyika, J., et al. (2022). The AI Index 2022 Annual Report (Stanford Institute for Human Centered AI).

20. Watkins, E.A., McKenna, M., and Chen, J. (2022). The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. Preprint at arXiv. 220209519. http://arxiv.org/abs/2202.09519.

21. Barocas, S., and Selbst, A.D. (2016). Big data's disparate impact. Calif. Law Rev. 104–671. https://www.ssrn.com/abstract=2477899.

22. (2021). High Level Group on Non-discrimination, Equality and Diversity, Subgroup on Equality Data. Guidance Note on the Collection and Use of Equality Data Based on Racial or Ethnic Origin (European Commission), p. 57. https://doi.org/10.2838/06180.

23. Moss, E., and Metcalf, J. (2022). The Social Life of Algorithmic Harms (Data & Society). https://datasociety.net/announcements/2021/10/28/the-social-life-of-algorithmic-harms/.

24. Sloane, M. (2021). The Algorithmic Auditing Trap (Medium OneZero). https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.

25. Rhea, A.K., Markey, K., D'Arinzo, L., Schellmann, H., Sloane, M., Squires, P., Khan, F.A., and Stoyanovich, J. (2022). An external stability audit framework to test the validity of personality prediction in AI hiring. Preprint at arXiv. http://arxiv.org/abs/2201.09151.

26. Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 501–512. http://arxiv.org/abs/1912.03593.

27. Moss, E.D. (2019). Translation Tutorial: toward a theory of race for fairness in machine learning. In FAT* Conference (ACM).

28. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Conference on Fairness, Accountability, and Transparency (FAT* '20), p. 12.

29. Sloane, M., and Moss, E. (2022). Introducing a Practice-Based Compliance Framework for Addressing New Regulatory Challenges in the AI Field (TechReg Chronicle), p. 9. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4060262.

30. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., and Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (National Institute of Standards and Technology). https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

31. IEEE Standards Association (2022). IEEE 7000™ Projects (IEEE Ethics in Action). https://ethicsinaction.ieee.org/p7000/.

32. Government of Canada (2020). Algorithmic Impact Assessment (AIA) (Responsible Use of Artificial Intelligence). https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

33. US CIO Council (2022). Algorithmic Impact Assessment. https://www.cio.gov/aia-eia-js/#/.

About the authors

Serena Oduro is a research analyst on the policy team at the Data & Society Research Institute. Serena's experience in genocide studies and technology ethics have ignited her passion for race-conscious technology policy. Previously, Serena was the 2020–2021 Technology Equity Fellow at the Greenlining Institute.

Emanuel Moss, PhD, is a socio-technical systems researcher at Intel Labs. He is an anthropologist who studies machine learning practices, knowledge production, and the social implications of science and technology. Additionally, he works on developing methods for algorithmic accountability that serve the public interest.

Jacob Metcalf, PhD, is the director of the AI on the Ground Initiative at the Data & Society Research Institute. He is a technology ethicist and studies ethics and accountability practices in data science research and the AI/ML industry. He is also a co-principal investigator on the National Science Foundation (NSF)-funded PERVADE Project and co-founder of the ethics consultancy Ethical Resolve.