Algorithmic Impact Assessments at Scale: Practitioners' Challenges and Needs

Amar Ashar, Karim Ginena, Maria Cipollone, Renata Barreto, and Henriette Cramer

Abstract. Algorithmic Impact Assessments (AIAs) are often suggested as a tool to help identify and evaluate actual and potential harms of algorithmic systems. While the existing literature on AIAs provides a valuable foundation, critical understanding gaps remain, including the lived experiences of practitioners who implement assessments and a lack of standardization across industry. Such gaps pose significant risks to the usefulness of assessments in the responsible development of algorithmic systems. By conducting 107 assessments with practitioners who build personalization, recommendation, and subscription systems at a large online audio-streaming platform and 8 semi-structured stakeholder interviews, we attempt to bridge this gap by identifying practitioners' challenges when applying AIAs that might hinder their effectiveness. The paper analyzes whether harms described in the literature related to machine learning and recommendation systems are similar to the concerns practitioners have. We find that the challenges practitioners encounter can be grouped into three categories: technical and methods, infrastructure and operations, and resourcing and prioritization. We also describe ways for teams to more effectively mitigate concerns. This paper helps bridge gaps between the theory and practice of AIAs, advances understanding of the potential harms of algorithmic systems, and informs assessment practices to serve their intended purpose.

1 Introduction

Algorithmic systems, ranging from Artificial Intelligence (AI)-based systems to heuristics, play an increasingly prominent role in advancing creativity, discovery, and culture (Wang and Siau 2019; Mökander et al. 2021; Suleyman 2023; Born et al. 2021). For instance, on digital entertainment platforms, automated systems often filter through millions of

© Stanford Internet Observatory, 2024
doi:10.54501/jots.v2i4.206

pieces of content in order to present a short list of recommendations personalized to users. (Stray et al. 2023)

At the same time, algorithmic systems have the potential to fail and cause significant harm to people and society (O'Neil 2017; Buolamwini and Gebru 2018; Noble 2020; Rubel, Castro, and Pham 2021; Weidinger et al. 2021). In anticipation of these harms, organizations need to carefully set up governance mechanisms that enable them to prudently develop and monitor algorithms to mitigate against these failures (Tabassi 2023). Algorithmic Impact Assessments (AIAs) are tools that are often suggested for systematic evaluation of potential societal impacts and are a technique often included within a suite of tools used for algorithmic auditing. Policymakers, civil society groups, academics, journalists, and industry leaders consider AIAs to be instruments that assist in the process of building or evaluating algorithmic systems that aim to foster trust, mitigate harm, and maximize benefits (Sandvig et al. 2014; Ada Lovelace Institute 2020; Brown, Davidovic, and Hasan 2021).

Algorithmic Impact Assessments conducted within organizations by practitioners, often referred to as first-party audits, are characterized as a proactive accountability mechanism for addressing potential algorithmic harms, including issues like bias, discrimination, and fairness, and these instruments have gained increased attention in recent years (Raji et al. 2020; Reisman et al. 2018; Moss et al. 2021). Second- and third-party audits are conducted with or by independent groups, including journalists, academics, or consulting firms. Examples of algorithmic audits and assessments include Microsoft's Responsible AI Assessment for GitHub's Copilot, a collaboration between Northeastern University and talent identification firm Pymetrics to audit a candidate screening tool, and an independent investigation of the TikTok algorithm by the *Wall Street Journal* (Microsoft 2024; Wilson et al. 2021; WSJ Staff 2021). AIAs are grounded in a long tradition of public and private sector use. For instance, impact assessments play critical roles across several domains, including privacy (Clarke 2009), data protection (Ivanova 2020), human rights (Kemp and Vanclay 2013), and environmental protection (Glasson and Therivel 2019). However, considering the widespread suggested use, there is a surprising lack of research on efficacy in scaled practice.

A recent review of the scholarship on AIAs found that while there is some convergence in the field highlighting their importance, there remains a lack of agreement on content, structure, and implementation (Stahl et al. 2023). Metcalf et al. (2021) warned against AIAs becoming a perfunctory exercise that "does not account for the harms algorithmic systems can engender in practice." Watkins et al. (2021) presented six concrete observations on how AIAs constitute accountability, and highlighted the need for empirical research on how AIAs intersect with the social processes of particular contexts. Stahl et al. (2023) called on researchers to examine the "organizational embedding" of AIAs to help us understand how they are applied, emphasizing the importance of studying AIAs in situ. This study brings attention to the complexities involved in operationalizing AIAs as

governance tools. We attempt to shed light on the practical challenges of implementing assessments and to derive novel findings that are beneficial for organizations.

This manuscript offers an analysis of 107 Algorithmic Impact Assessments conducted on AI- and non-AI-based systems and 8 semi-structured stakeholder interviews at a large streaming media platform, and seeks to understand how existing harms frameworks proposed by researchers map onto practitioners' experiences. The platform offers music, podcasts, and other personalized content, and operates in over 180 countries worldwide. By conducting assessments with practitioners who build personalization, recommendation, and subscription systems and facilitating interviews within the same organization about assessment implementation, we attempt to bridge the gap between industry and academic research by identifying practitioners' challenges when applying AIAs that might stand in the way of their effectiveness.

Our analysis finds that the challenges practitioners encounter can be grouped into three categories: (1) technical and methods, including product team questions about how to evaluate for fairness, bias, and adequate representation; (2) infrastructure and operations, reflecting how teams within large-scale organizations need to address both technical debt and organizational responsibilities; and (3) resourcing and prioritization, including organizational challenges requiring programmatic structures and guidance for prioritizing algorithmic and AI responsibility efforts. We also describe ways for teams to more effectively mitigate concerns, including through the development of applied guidance and governance, engagement with both internal and external stakeholders, and applied product operational mitigation such as tooling and metrics.

2 Background and related work on Algorithmic Impact Assessments

2.1 Endorsement of AIAs by governmental and international bodies

In recent years, there has been an increase of national and international bodies endorsing Algorithmic Impact Assessments as a means to encourage developers of automated systems to think critically and methodically about these systems and to document their decisions and improve accountability (Selbst 2021). Interdisciplinary groups, including the US National Institute for Standards and Technology (NIST) and the IEEE, are working to foster nascent AI responsibility standards. NIST states that "risks from AI-based technology can be bigger than an enterprise, span organizations, and lead to societal impacts" (Tabassi 2023). Accordingly, it recommends that organizations govern, map, measure, and manage these risks. AIAs are one such tool that NIST recommends for "assessing and evaluating requirements for AI system accountability, combating harmful bias, examining impacts of AI systems, product safety, liability, and security, among others." Furthermore, the "H.R.5628 - Algorithmic Accountability Act of 2023" (2023)

requires impact assessments on automated decision systems and augmented critical decision processes as a means "to document known harm, shortcoming, failure case, or material negative impact on consumers," among other requirements.

Similarly, the International Organization for Standardization (ISO), an independent nongovernmental global organization that develops standards, stipulates in ISO/IEC 42001 (2023) "the need for organizations to establish an impact assessment process to determine the impact an AI system has on individuals and societies." ISO is currently developing a standard (ISO/IEC 42002 2024) on AI impact assessments, and has addressed the topic more broadly in previous standards (e.g., ISO/IEC 23894 (2023), ISO/IEC 38507 (2022), and ISO/IEC 42001 (2023)).

More recently, the European Commission, according to the Council of European Union EU AI Act (2021), requires high-risk AI systems to undergo an impact assessment to determine if they conform to the Act's requirements. This assessment is needed prior to launching systems in the EU market or if they are substantially modified after they were launched.

These developments and endorsements of AIAs point to the important role this tool plays in the rapidly evolving governance landscape. Despite the growing body of literature on assessments, the significant lack of research exploring the real-world experiences of practitioners conducting AIAs presents a critical challenge that hinders their full potential. Gleaning insights from practitioners' experiences is vital to enhance the applicability of AIAs, and, in turn, leads to more practical and effective governance. Practitioners implementing AIAs witness rapid technological changes in real time, and their experiences offer unique and valuable insights into how they account for and adapt to the complexities and unpredictability they encounter. Bridging the gap between theory and practice is imperative for refining our processes, building practically viable AIAs, and addressing blind spots.

2.2 How industry seeks to operationalize AIAs

Although few standards exist for what type of information should be collected and assessed using AIAs, technology industry researchers and practitioners have proposed different forms of assessment and auditing tools to foster transparency, accountability, and efficiency. Raji et al. (2020), with colleagues from Google, were among the first to describe an "End-to-End Framework for Internal Algorithmic Auditing," which focuses on evaluation of each stage of the machine learning lifecycle. Microsoft uses an extensive AI Impact Assessment template and an associated AI Impact Assessment Guide that offers practical information for teams building AI-based products to manage risks as part of a suite of tools and practices for responsible AI. (Microsoft 2022)

Groups like Telefónica, Spotify, the BBC, and the Ada Lovelace Foundation also describe their use of Algorithmic Impact Assessments to provide practitioners with training,

centralized guidance, and actionable tools to identify and address issues brought to light during an assessment process (Telefónica 2021; Ashar and Cramer 2022; BBC 2021; Ada Lovelace Institute 2020). Algorithmic audits on specific systems, such as Twitter's (now X) image cropping algorithms or independent work on Amazon's product recommendations promoting health misinformation, and more systemic examinations like Airbnb's Project Lighthouse, also serve as examples of the types of deeper-dive audits that assessments may catalyze (Chowdry 2021; Juneja and Mitra 2021; AirBnB 2022).

Researchers have proposed a set of tools, increasingly adopted by practitioners, which includes model cards and system cards. These tools help document models/systems and their intended use, provide information about data and technical design, and often describe benchmarks for safety and performance. Google and Hugging Face, for example, publicly describe their use of model cards and their utility for model transparency (Mitchell et al. 2019; Hugging Face 2023). Meta and OpenAI employ system cards, which describe how groups of models and other types of systems interplay to produce a certain outcome, reflecting the larger interdependent ecosystem of technology and enabling systems to be more explainable and auditable (Alsallakh et al. 2022; OpenAI 2023).

2.3 Frameworks for conceptualizing algorithmic harms

Algorithmic Impact Assessments aim to identify and reduce potential algorithmic harms. Such harms are a significant cause for concern with automated systems, especially as these systems have become increasingly integrated into our daily lives (Buolamwini and Gebru 2018; Perez 2019). But harms can vary, and can extend beyond bias to cover other types of harms such as invasion of privacy and misinformation.

Several frameworks have been proposed in the literature to classify algorithmic harms. In one such framework, the primary framework that influenced this AIA study, Crawford (2017) distinguishes between two types of harms caused: Harms of allocation, or unfairly assigned opportunities or resources due to algorithmic decisions, and harms of representation, or stereotypical depictions of groups of people that further marginalize them in society. Crawford notes the former is immediate, easily quantifiable, discrete, and transactional, while the latter is long-term, difficult to formalize, diffuse, and cultural. Shelby et al. (2023) add three other types of harms to Crawford's framework: quality of service, interpersonal, and social systems harms. Quality of service harms are disparities in performance whereby automated systems underperform for certain groups of people. Interpersonal harms are characterized by adverse impact on relationships between people. Social system harms refer to adverse macro-level effects of algorithmic and automated systems on groups of people.

Algorithmic Impact Assessments may focus on specific types of harms, and categorize and translate harm categories in ways that fit the particular context of use. Translation of

both academic and legal frameworks might be necessary within industrial organizations. Such translation may depend on what an application is used for, stakeholder concerns, and specific technologies in use, and may require terminology more familiar to product teams and stakeholders. To a certain extent, the harm categories used in practice in AIAs are a matter of definition and practicality. For example, in this specific study, harms related to quality of service (e.g., whether different categories of listeners get similarly relevant results when audio content is recommended to them), are categorized as a subtype of harm of allocation; i.e., a level of service is seen as a resource.

In certain cases it might be appropriate to start with a framework tailored to the particular type of machine learning used. For example, Weidinger et al. (2021) focuses on language models and categorizes potential harms into six categories: (1) discrimination, exclusion, and toxicity; (2) information hazards; (3) misinformation harms; (4) malicious uses; (5) human-computer interaction harms; and (6) automation, access, and environmental harms. In other cases, it might be prudent to start with guides for harms and mitigation in a specific sector. Leslie (2019) classifies harms into six categories while focusing on usage of AI systems in the public sector: (1) bias and discrimination; (2) denial of individual autonomy, recourse, and rights; (3) non-transparent, unexplainable, or unjustifiable outcomes; (4) invasions of privacy; (5) isolation and disintegration of social connection; and (6) unreliable, unsafe, or poor-quality outcomes. Lee et al. (2024) in turn further present a taxonomy of 12 subtypes of AI privacy risks, which might be appropriate if the focus of an AIA is on privacy risks created or exacerbated by Artificial Intelligence-based systems.

Measurement challenges are also noted across frameworks as a limitation and as a need for further research. In some cases, this is even the basis of harm classification itself; Hoffmann and Frase (2023) classify harm into tangible (i.e., observable, verifiable, and definitive) and intangible (i.e., can't be directly observed) harms. Tangible harms are broken down into harms to physical health or safety, infrastructure or the environment, and financial loss and damage to property. Intangible harms are broken down into detrimental content, differential treatment, and harms to privacy, human/civil rights, and democratic norms. Frameworks and operationalization of algorithmic responsibility through AIAs can, however, also reference methods to measure abstract concepts such as algorithmic fairness; implemented approaches may differ, and fairness measurement is often dependent on contextual factors like the intended use, outcome of the system, and data availability (Smith, Beattie, and Cramer 2023). In this study, one function of the AIA beyond recommending mitigation approaches for immediate harms (e.g., using particular types of data or using specific content systems to enable standardization), was to also gather information about what types of measurement teams themselves found effective.

All of these varied frameworks illustrate how the harms of automated systems can be examined and conceptualized from different perspectives. Indeed, there is no shortage

of frameworks in the literature on the topic. If anything, this illustrates the importance of understanding harms using an interdisciplinary lens and working with a diverse set of stakeholders (e.g., technical, policy, legal, etc.) to set up governance mechanisms that curtail the chances of these harms occurring. However, the abundance of frameworks also shows the practical need to make a decision within an organization about which framework should be used or adapted to form the basis of an organizational-wide understanding of what harms to address. For AIA implementation, this means that a situationally appropriate framework or specific harms may need to be first selected and adapted to be able to guide model developers on what harms to look out for. Practioners using AIAs can accordingly provide feedback on whether a chosen framework actually covers potential concerns about harms that teams or other stakeholders have encountered or considered.

3 Organizational context

By analyzing first-party AIAs and conducting interviews at an audio-focused streaming media platform, this study assesses the challenges encountered when instrumenting AIAs in large organizations that deploy hundreds of models. We believe it is crucial to understand the organizational context in which AIAs are used for effective implementation and the resulting challenges encountered.

In 2021, the platform that is the focus of this study created a centralized set of requirements for teams that develop algorithmic systems. This was designed to help practitioners across the organization more efficiently and consistently assess and address potential algorithmic harms, prepare for external legal and governmental requirements, and improve recommendation systems. While algorithmic systems enable opportunities for audio creators and listeners alike, policy teams working in collaboration with product groups want to ensure that the development, deployment, and impact of products are as safe and equitable as possible. Considering the scale of algorithmic development, organizations with a large user base often have dozens or hundreds of models operating in parallel, so coupling centralized guidance for responsibility with decentralized ownership and development of systems is important. At the time of this study, the primary focus of the platform's offerings included music and podcasts, and assessments covered AI or algorithmic systems that either directly or indirectly played a role in serving that content to users or supporting business functions.

The centralized requirements for algorithmic systems were operationalized through an assessment rollout process that included policy documentation; organization-wide training for personalization, recommendation, and subscription product areas; and standard operating protocols and procedures. The impact assessment included a template/survey of questions generated through an academic and industry literature review of current practices. For the AIAs conducted at this streaming platform, representatives and prod-

uct teams were trained using Crawford (2017)'s framework of high-level categories, and data from assessments were used to advance internal frameworks for harm specific to the product and context of the organization.

Algorithmic Impact Assessments on platform systems were designed to enable teams to (1) check baseline compliance with central requirements such as infrastructure tool usage and data restrictions; (2) proactively find potential algorithmic issues before they significantly impact users or creators; (3) adequately plan and address problems identified; (4) manage risk and harms for highest-priority surfaces and users; and (5) inform product and platform strategy. Assessments help teams to address issues by identifying gaps in coverage and resources; alerting researchers to where existing or new methods, tools, data, and expertise are needed; and prioritizing work ahead. The AIA template included more than two dozen questions for teams collecting information about their systems, data usage, content recommendation design and policy, equitable outcomes and mitigation of harm, and ownership/lineage. AIAs were not intended to be comprehensive technical source code or model output audits for each model, which are formats often used for algorithmic assessment. Instead, AIAs in this context call attention to where more comprehensive investigation and work may be needed, such as product-specific guidance, deeper technical assessment, or external review. A limitation of current work and area of future study is how assessments can directly engage with stakeholders that may have experienced specific harms.

4 Methods

4.1 Assessment template

Practitioners completed a total of 107 AIAs for different automated decision systems across groups that built personalization, recommendation, and subscription systems. The number of AIAs on systems reflects a significant portion of personalization and recommendation systems at the company that were identified as high priority by product teams.

Information from the AIAs analyzed in this study included responses to the following questions:

- Have you conducted quantitative evaluations or have quantitative measurements of where there may be unequal outcomes for creators or listeners in your system? If so, please describe the test and results. If not, please describe the tests that are necessary, and offer a sense of when you plan to conduct them.
- Have any changes been made to address potential harms, for example after evaluations? If not, what could be done?

- Do you have a sense of the threshold for measuring and correcting impact to creators/listeners based on the aspects listed under the 'equitable outcomes' section of the policy in the system, especially as it relates to protected groups? Or how you might determine that threshold?
- What are realistic worst case scenarios in terms of how errors might impact society, individuals, and stakeholders? How confident are you in the decisions output by your algorithmic system to not exacerbate or perpetuate harms?
- Looking at your system holistically, what are potential risks in the system from your perspective? In particular, where might the performance vary?
- What are the potential sources of risk (e.g., training data, feature choice, model development, design choices.) If relevant, does your training data represent the diversity of your user base?
- What are the potential downstream effects you anticipate or have identified?
- Are there any obstacles or challenges (e.g. method questions, technical challenges or unavailable data) to test this system?
- If you could have external advice or external auditing of your system, what would your top wishes be and what would be most helpful?

Additional questions on the assessment, not included in this analysis, asked teams to document the function and intended use of models, dependencies on upstream and downstream systems, compliance with internal safety policies, and the teams' adherence to algorithmic responsibility best practices. Note that the assessment questions listed here and used for analysis were slightly edited to remove organization-specific references.

4.2 Scaffolding assessment implementation and analysis methods

Drawing on centralized materials, templates, and training described above, product teams were then assigned a trained representative for algorithmic responsibility. Representatives interfaced with product teams and helped product managers, engineers, or product leads complete the AIA process by operating as assessment reviewers, as well as implementers who created structured roadmaps that identified issue areas from assessments, planned short and long-term work ahead, and informed overall company guidance and strategy.

For the purpose of this research, we focused our analysis on questions about potential risks to individuals and society, challenges in conducting AIAs and testing systems, and team requests for support. Using thematic coding, we analyzed the qualitative data provided in a subset of fields across all AIAs responses, extracted common and salient themes, and summarized key findings.

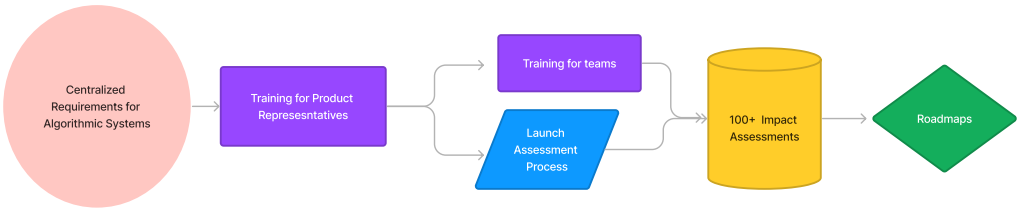

Figure 1: Algorithmic Impact Assessment Process.

4.3 Interviews

We conducted eight semi-structured user interviews with three types of internal users: engineers; product area leaders; and algorithmic responsibility liaisons (product representatives who interfaced with a central team). The goals of the research were to (1) evaluate participants' experiences with AIAs, understand how to improve upon them, and learn how to better leverage their equity-minded motivations in more tactical ways; (2) examine the impact of one's role on the algorithmic auditing process and how each contributes to the motivation to engage in responsibility processes and work; and (3) uncover tactics that contribute to the work and barriers to it. Previous qualitative research within the organization helped establish the centralized requirements, which governed the AIA process that was implemented at a greater scale across teams.

An experienced user researcher conducted interviews and asked questions from an interview discussion guide drafted in line with the research goals above. Prior to beginning the interview, participants were assured that the information they shared would be anonymized to make improvements in the AIA process and contribute to an academic understanding of the process. Interviewees were encouraged to be honest so as to drive the most positive changes, and were strongly encouraged to share as much as possible, including product materials and discussions, to demonstrate their answers. Before concluding, the interviewer asked participants if they had any questions and addressed them. They also previewed next steps, provided them with their contact information to offer a space for follow-up thoughts, and thanked them for their time. As participants were internal employees, no additional financial compensation was provided.

Findings from the analysis of AIAs below reflect issues raised by product teams within the context of building and deploying AI and algorithmic systems; findings from the semi-structured interviews describe lessons learned from the process of implementing AIAs at scale and what gaps and opportunities practitioners shared to improve organizational processes and support.

5 Findings: Examining potential and existing harms Through practice

5.1 Practitioner-discussed harms

Drawing on the frameworks described and additional literature in algorithmic auditing, as a starting point teams were provided with guidance and examples of the types of algorithmic harms researchers and other industry practitioners have encountered when evaluating automated systems. Within the AIAs, practitioners were asked to identify whether any potential or current harms existed from their perspective and to expand on whether their systems or models may present harms not yet described in the resources provided. Given the audio-focused nature of the platform, many teams assessed user-facing feature systems as well as the underlying systems that power multiple types of music or podcast content.

The first category of harms we observed from data based on the written assessments was potential harms of allocation. The potential for opportunity loss was most prominent in recommendation systems due to system exclusions, ranking-related exclusions, and editorial and third-party exclusions. For instance, practitioners noted that a model that produced algorithmically curated compilations of an artist's catalog only generated such compilations if an artist has a minimum set number of tracks, among other heuristics.

System exclusions can occur when the design or implementation of an algorithmic system excludes certain groups. For instance, a system might redirect traffic away from a musical artist because the model excludes certain genres based on the interest or past listening behaviors of users, such as an algorithmically generated set of tracks informed by a user's age in order to create a nostalgic playlist. Such actions can lead to a lack of representation or access to opportunities for the excluded groups or users who are searching for content that may not suit listener preference or behaviors.

Ranking-related exclusions, another means for opportunity loss, refers to the practice of ranking or categorizing creators based on certain criteria, for instance, making it difficult for creators with fewer listeners to show up in contexts where popular listening is highlighted. Popularity bias could have a direct effect on newer or more independent creators, impacting their chances to succeed or climb the ladder for streaming or further economic success. For example, certain assessments noted and researchers have found that balancing the relevance of recommendations to users and fairly representing creators is important to mitigate the effects of superstar economics (Mehrotra et al. 2018).

Editorial and third-party exclusions, another avenue for opportunity loss, can occur when choices made by editors, such as excluding a category of genres, are codified in algorithms, leading to a skewed representation of categories. Such a skew could also

occur through third-party systems that act as components of models, such as those implemented through external vendors.

Beyond opportunity loss, harms of allocation may also result in economic loss. Potential additional harms described by respondents included the potential for loss of revenue for creators, wrongful denial of services to users, and disparity in user pricing or trial offers. For instance, if algorithmic systems only recommend popular creators, then fewer creators are able to make a living, leading to further exclusion of a large long tail of creators. Wrongful denial occurs when genuine groups of users have their subscriptions terminated or their payments incorrectly denied because of system errors. Finally, in the context of digital platforms, economic loss may also occur when the audience for a product or service reflects certain demographic characteristics and ends up paying more or less than others for the same service. There is also potential for certain features or feature tests to be allocated to users with advanced or expensive technology (e.g., mobile devices with most up-to-date hardware/software), contributing to economic inequality.

A second category respondents described in written assessments reflected harms of representation, which can manifest in a number of ways. First are cultural harms, where a system, for example, promotes a podcast in one region to users of another region where it may be considered inappropriate, irrelevant, or unwanted. Second, underrepresentation may mean that users do not see content in their language, leaving those listeners feeling unimportant or distanced. Third, with loss of agency, the system presents inadequate image choices for artists to pick from, such as when artists become associated with other artists they may not necessarily identify with genre-wise, or an editor chooses art that may not be fully representative of a content collection. This can result in creators or listeners feeling that their experiences and identities are not authentically portrayed or valued. Fourth, stereotyping or demeaning groups could be the result of descriptive terms being inaccurate, misleading, or hurtful representations, or users from particular groups could be mistakenly flagged as potential account abusers and asked to go through a re-verification procedure when they should not have had to do so.

A third category of harms identified in our data from assessments highlighted quality of service harms. We found that a specific system's performance could potentially vary by country/market, language, genre, topic, demographics, and chosen payment method. For instance, users in emerging markets, where limited user data is available to train algorithmic systems, may have received less relevant recommendations during the launch of the service than users from large markets, where more data exists for training purposes. Another example from our data is fraud detection systems that take into account registration information as part of their inputs and may unintentionally impact certain accounts because of how quickly fraudsters' practices evolve. A third example is potential underperformance of models for users searching for niche genres or those who use non-English descriptions in the process of searching. Certain models that are

primarily trained on popular genre descriptions may not perform as well for non-popular descriptive terms.

Although assessments took place on a system level, practitioners also reported that the complex interdependence of different types of systems to produce a user-facing outcome, including those upstream and downstream of their services, may create gaps in their ability to anticipate harms when they do not always have a full end-to-end picture.

5.2 Remediations

Addressing the range of potential and existing harms described by product teams in the assessments required both substantive guidance and organizational support. We were especially interested in mitigations that teams had already applied that may not have been previously described in documentation—or in easily digestible external literature—that could serve as examples for other teams. Product teams provided a number of cases in the assessments, including already performed or planned work.

Teams described substantive areas for remediations in response to the question in the assessment about changes that have or could potentially be made, including further testing and monitoring of harms primarily focused on demographic impact, support to develop fairness or responsibility metrics, feature or data removal or changes in machine learning models, and model retraining. In addition to model-focused areas, practitioners suggested that product and editorial interventions such as diversifying or changing how pools of content were chosen, filtering content based on policy measures or product goals, or editorial curation of content could assist with remediation.

Feedback about organizational remediations were primarily reflected within the semi-structured interviews. Practitioners expressed a desire to see substantive changes prioritized on product roadmaps, and for product leadership to communicate not only the importance of addressing responsibility concerns, but also how policies and governance mechanisms need to evolve based on product feedback. It is noteworthy that for a number of these mitigations—such as editorial interventions—in-depth domain and application knowledge are necessary, even to understand the answers provided in the AIAs. This illustrates the importance of preserving organizational knowledge about prior solutions, stakeholder reports, and incidental findings by teams themselves. It also underscores the need for researchers to deepen understanding of the tacit knowledge engineering and product teams have and their (practice-proven) techniques.

6 Mapping found harms and processes onto frameworks

Our findings did not neatly map to any single proposed framework on algorithmic harms, but suggest that such frameworks are a helpful starting point and that harms frameworks for audio-focused media may need further development. They did, however, most

closely resemble the conceptualizations of Crawford (2017) and Shelby et al. (2023), underscoring the importance of organizations not indiscriminately adopting frameworks found in the literature, but examining and contextualizing harms to get a more accurate understanding of algorithmic systems data, design, and intended outcomes.

Our findings also illustrate how new manifestations of harms can be unearthed under the broad categories (e.g., harms of allocation), enriching the literature and building a stronger understanding of the root causes of these harms. A lesson learned from this effort is that academic frameworks and terminology, when shared with product teams even with training and context, can often seem too vague or broad to spur teams to further action—whether that means discovery, testing, or mitigation. Practitioners themselves are a rich source of feedback for where internal policy or guidance may need finer requirements, including new types of research methods, support for evaluation, priority-setting from leadership, or suggestions for interventions or mitigations to test.

Teams also need to act quickly during product development and do not often have the privilege of time to wait for research teams to evaluate new methods to identify or mitigate algorithmic harm, for fairness or safety metrics to be established at scale before making changes to their models, for product leadership to articulate each change as a priority, or for recommendations to be instantiated into policy and standard operating procedures. Investing further in self-service tooling, establishing practices incorporated into engineering protocols, supporting safety by design, and combining practitioners' real-world experience with broader field research may help to address potential gaps between existing frameworks and the types of harms that may appear on streaming media platforms. No one framework may comprehensively cover all use cases, nor was the objective of this study to map to a singular framework; however, for researchers exploring the range of harms emerging from AI and algorithmic systems, future iterations of frameworks may benefit from more specific suggestions that are applicable to specific domains. For example, algorithmic harms emerging from AI used in criminal justice contexts may be distinct from those users consuming media from AI-based recommendation systems, even if the overarching categories of harms of representation and allocation are shared. Studying harms related to culture in algorithmic contexts presents an opportunity for researchers to empirically understand where there is impact in culture, streaming, and creator-focused economies, as well as inform development of broader harms frameworks.

7 Lessons from practitioners about AIA processes

While direct answers to open product questions are not always available, neither in the research literature nor in commercially available tooling, the Algorithmic Impact Assessment process ultimately helped to:

- Establish formal workstreams that led to improvements in recommendations and search.
- Improve data availability to better track platform impact.
- Reduce data use by minimizing data usage to what is actually necessary for better product outcomes for listeners and/or creators, beyond data privacy compliance alone.
- Help teams implement additional safety mechanisms to avoid unintended amplification and ensure consistency.
- Inform efforts to address dangerous, deceptive, and sensitive content, as described in the platform's rules.

However, AIAs do not always deliver organizational momentum and solutions, nor visibility of their impact. In the following section, we transition from conceptualizing harms to building an understanding of the experiences of practitioners who undertook AIAs. The objective here is to uncover the pain points of the process in an attempt to highlight areas that require attention and improvement. The first section addresses the challenges and obstacles that participants experienced and the following section describes where they reported needing the most help.

7.1 Practitioner challenges

7.1.1 Technical and methods

Technical- and methods-related challenges were often specific to the systems a team owned. Some of the reported challenges teams encountered include how to evaluate model fairness, how to determine if model underperformance in certain regions of the world is due to implicit data bias including underrepresentation, limitations on the types of analysis that can be run due to the speed with which a system operates (e.g., composes content within 100 ms of request), and scaling challenges, among others. Teams also wanted an agreed-upon framework for understanding potential harms in order to be able to interpret their results. A final set of challenges reported in this first theme related to metrics, baselines, and thresholds. Teams faced difficulties measuring the long-term impact of system decisions, or found that they lacked adequate metrics to measure algorithmic harm, including baselines to measure against and thresholds on which to act.

7.1.2 Infrastructure and operations

Technical infrastructure- and operations-related challenges were common. Product groups reported a key problem with legacy systems handed down to teams due to reorganization, which are expected in organizations that have existed for many years. Legacy systems often act as components of a larger product experience, and can pose

a challenge due to the complexity of where they fit in that product chain, the age or authorship of the code used to build them, and a relative lack of working knowledge associated with these systems within teams. Another obstacle described by teams was the availability of data to run the necessary tests and concerns around the quality of such data. A final challenge expressed by product teams was the demarcation of responsibilities between teams for systems that take data from upstream systems or that are a data source for downstream systems. The independence of teams and interdependence of systems often made it difficult for downstream teams to influence the prioritization of algorithmic issues of those upstream.

7.1.3 Resourcing and prioritization

Organizational resourcing- and prioritization-related challenges affected teams' ability to act quickly. Teams often reported a lack of time and engineering resources required to run tests based on centralized guidance provided to them. They often found it difficult to prioritize work self-identified in assessments against competing commitments, especially until responsibility work was encoded in organizational-wide priorities or efforts. Despite clear interest from participants to dig deeper into these issues, the need to prioritize made it difficult for them to give these activities any more time or effort because that could come at the expense of other product needs. Many teams leaned on product representatives, who had additional domain knowledge and training, in order to move product roadmaps for responsibility forward. AIAs also underscored the need for continued policy development, frameworks for this specific organizational context, and decision-making.

7.2 Practitioner needs

7.2.1 Applied guidance and governance

Teams reported looking for more training opportunities to understand how to identify, monitor, and potentially mitigate potential harms created by user-facing systems or underlying models, particularly in the context of recommendations. They reported a need to understand how to minimize bias throughout the entire product lifecycle and how to assess the overall trustworthiness of systems. They also asked for case studies of internal teams that have done things right to understand what their "golden path" looked like and the sort of time and resources that might be needed. Teams also wanted playbooks to provide detailed guidance in order to take concrete next steps. Product developers expressed the need for more governance frameworks, including how they should be prioritizing work and who should be making decisions, and more consistent guidance on how to report issues and progress. Examples teams described included having a scorecard to prioritize the work on algorithmic harm, using model or system cards for more consistent reporting, tooling for automated monitoring and alerts, and support earlier in product life cycles to design for responsibility and safety.

7.2.2 Internal and external engagement

Practitioners proposed several working arrangements to address algorithmic harms more effectively. One suggestion included an embedded model, where a researcher specialized in algorithmic harms assisted them with measuring, evaluating, and improving the system. Another recommendation proposed a consultation service provided by a horizontal team that would provide necessary expertise to alleviate some of the team's load. The nature of help needed varied by team. For example, some asked for the horizontal team to do "external analysis" to complement their own analysis, while others asked for more holistic auditing of their products to identify areas of concern in order to be able to measure their impact. Beyond continuous learning opportunities, teams expressed interest in knowledge- and resource-sharing opportunities. They also reported a willingness to collaborate with other companies tackling biases in recommender systems as a form of building support across the tech sector on these issues. Teams expressed an interest in working with outside experts, suggesting that the first-party assessments conducted in this study may be complemented by second- and third-party assessments conducted by independent experts.

7.2.3 Operational mitigations

Practitioners asked for support to operationalize remediation, requesting additional methods for evaluating systems, acceptable thresholds for fairness metrics, and dashboards for monitoring metrics across different user slices (e.g., metrics for subsets of creators or users in certain demographics). Teams thought such thresholds would be useful because they could be alerted if they crossed them. Teams also needed guidance on the level of rigor that needed to be applied to internal systems vs. user-facing products. In addition, they asked to see hypothetical scenarios where metrics were adversely impacted by bias in protected classes (e.g., gender), to build a more concrete understanding of how damaging this effect is on users. They also expressed interest in interacting with a simulation to get a real feel for impact. Centralized teams could also play a greater role in translating data and metrics into insights or recommendations for product teams.

Appendix A summarizes the challenges and needs of practitioners with AIAs.

7.3 Driving intrinsic motivation for algorithmic responsibility work

However, this project and research revealed a core learning: the need to gear assessment processes toward practitioner priorities.

We learned from our research that when practitioners felt that they had to execute AIAs and enforce the corresponding algorithmic policies and centralized guidance because "it's the right thing to do," they were not always intrinsically motivated to do so—especially when processes are seen as overhead. For example, in several interviews participants

noted that the bigger the gap between high-level guidance on algorithmic responsibility and the practical needs of teams, the more room that leaves for ambiguity to develop, and the less invested teams might become because of uncertainty of actionability of the work. In contrast, applied guidance helped employees and teams create achievable tasks that sustained their intrinsic motivation. Another lesson was that employee recognition on topics related to algorithmic responsibility helped build employees' internal motivation and demonstrate the priority that this work carried in the eyes of leaders at the organization. Further, incorporating expectations on algorithmic responsibility as part of goal-planning and performance reviews not only helps organizations reap the rewards of these institutional mechanisms, but also encourages employees to direct attention to this line of work. These initiatives in tandem help build a supportive culture that nourishes the work on algorithmic responsibility and reinforces the intrinsic motivation of employees.

8 Discussion and Conclusion

This study bridges the gap between theoretical understanding and practical implementation of Algorithmic Impact Assessments, emphasizing the importance of understanding frameworks on algorithmic harms in light of the context and systems of organizations. It provides insights into the challenges that practitioners encounter in undertaking AIAs in an industry environment and areas where they need the most help. Our research underscores the importance of framing AIAs not as checklists, or as comprehensive documentation of "what could go wrong," but as dynamic tools that evolve in response to the unique challenges and needs of the practitioners who interact with them. AIAs must be implemented within the context of functional and implementable governance processes, and should be viewed as co-created organizational, social, and technical instruments. Operationalizing and exercising responsibility and safety within a complex environment requires a focus on people and processes in addition to the technical questions and problems to be addressed. Especially considering the recommendation to perform Algorithmic Impact Assessments from governmental, academic, civil society, and professional circles, it is critical to explore *how*, in addition to *why*, these processes can be performed effectively in actual and scaled contexts.

For resources that aim to inform practice, the focus should be on actionability rather than large-scale processes that could be perceived as bureaucratic. If it is unclear how the outcomes of large-scale assessments impact or change practitioners' actions, it is easy for resistance to develop to both the AIA process itself and the teams involved. Different types of stakeholders might react in distinct ways, and may benefit from varied formats of communication about the motivation and utility of assessment processes, beyond basic compliance. If the intention and purpose are unclear, the process needs to be adjusted. Note that the recommendation is not only to "educate," but rather to gear any similar process toward actionable findings, clear communication, and prioritized

order of mitigation work.

The findings of this paper contribute to the conversation on the responsible use of AI, algorithmic, and automated systems and emphasize the need for better resource allocation to ensure that Algorithmic Impact Assessments are integral to the responsible development and deployment of algorithms. Finally, this study serves as a reminder of the importance of ongoing, practice-informed research, and calls for stronger collaboration between academia and industry practitioners to ensure that AIAs are adapted to keep pace with the rapidly evolving landscape, ultimately leading us toward more equitable and transparent algorithmic systems.

References

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

AirBnB. 2022. *A Six-Year Update on Airbnb's Work to Fight Discrimination,* December 13, 2022. https://news.airbnb.com/sixyearadupdate/.

Alsallakh, Bilal, Adeel Cheema, Chavez Procope, David Adkins, Emily McReynolds, Erin Wang, Grace Pehl, Nekesha Green, and Polina Zvyagina. 2022. *System-Level Transparency of Machine Learning.* Technical report. Menlo Park, CA: Meta, February 22, 2022. https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/.

Ashar, Amar, and Henriette Cramer. 2022. *Lessons Learned from Algorithmic Impact Assessments in Practice.* Spotify Engineering, September. https://engineering.atspotify.com/2022/09/lessons-learned-from-algorithmic-impact-assessments-in-practice/.

BBC. 2021. *Responsible AI at the BBC: Our Machine Learning Engine Principles,* May. https://www.bbc.co.uk/rd/publications/responsible-ai-at-the-bbc-our-machine-learning-engine-principles.

Born, G., J. Morris, F. Diaz, and A. Anderson. 2021. *Artificial Intelligence, Music Recommendation, and the Curation Of Culture.* Schwartz Reisman Institute for Technology & Society, June 1, 2021. https://srinstitute.utoronto.ca/s/Born-Morris-etal-AI_Music_Recommendation_Culture.pdf.

Brown, S., J. Davidovic, and A. Hasan. 2021. "The Algorithm Audit: Scoring the Algorithms That Score Us." *Big Data & Society* 8, no. 1 (January 28, 2021): 205395. https://doi.org/10.1177/2053951720983865.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, February 24, 2018. https://proceedings.mlr.press/v81/buolamwini18a.html.

Chowdry, Ruman. 2021. *Sharing Learnings about Our Image Cropping Algorithm.* Twitter, May 19, 2021. https://blog.x.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

Clarke, Roger. 2009. "Privacy Impact Assessment: Its Origins and Development." *Computer Law & Security Review* 25 (2): 123–35. https://doi.org/10.1016/j.clsr.2009.02.002.

Council of European Union. 2021. *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), Council Regulation (EU) no 2021/0106(COD).* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

Crawford, K. 2017. "The Trouble with Bias Conference on Neural Information Processing Systems." https://www.youtube.com/watch?v=fMym_BKWQzk.

Glasson, John, and Riki Therivel. 2019. *Introduction to Environmental Impact Assessment.* Routledge, February 27, 2019. https://doi.org/10.4324/9780429470738.

"H.R.5628 - Algorithmic Accountability Act of 2023." 2023. Accessed January 22, 2024. https://www.congress.gov/bill/118th-congress/house-bill/5628/text.

Hoffmann, M., and H. Frase. 2023. *Adding Structure to AI Harm.* Center for Security / Emerging Technology, July. https://doi.org/10.51593/20230022.

Hugging Face. 2023. *Model Cards.* https://huggingface.co/docs/hub/en/model-cards.

ISO/IEC 23894, International Organization for Standardization. 2023. *ISO/IEC 23894:2023 — Information technology — Artificial intelligence — Guidance on risk management.* https://www.iso.org/standard/77304.html.

ISO/IEC 38507, International Organization for Standardization. 2022. *ISO/IEC 38507:2022 — Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations.* https://www.iso.org/standard/56641.html.

ISO/IEC 42001, International Organization for Standardization. 2023. *IEC 42001 Information Technology — Artificial Intelligence — Management System.* https://www.iso.org/obp/ui/en/#iso:std:iso-iec:42001:ed-1:v1:en.

ISO/IEC 42002, International Organization for Standardization. 2024. *IEC 42002 Information Technology — Artificial Intelligence — AI System Impact.* https://www.iso.org/standard/44545.html.

Ivanova, Yordanka. 2020. "The Data Protection Impact Assessment as a Tool to Enforce Non-discriminatory AI." In *Privacy Technologies and Policy: 8th Annual Privacy Forum, APF 2020, Lisbon, Portugal, October 22–23, 2020, Proceedings 8,* 3–24. Springer, Cham. https://doi.org/10.1007/978-3-030-55196-4_1.

Juneja, Prerna, and Tanushree Mitra. 2021. "Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 1–27. CHI '21. Yokohama, Japan: Association for Computing Machinery, May 21, 2021. https://doi.org/10.1145/3411764.3445250.

Kemp, D., and F. Vanclay. 2013. "Human Rights and Impact Assessment: Clarifying the Connections in Practice." *Impact Assessment and Project Appraisal* 31, no. 2 (May 22, 2013): 86–96. https://doi.org/10.1080/14615517.2013.782978.

Lee, Hao-Ping (Hank), Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. "Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks." In *Proceedings of the CHI Conference on Human Factors in Computing Systems,* 1–19. CHI '24. Honolulu, HI, USA: Association for Computing Machinery. https://doi.org/10.1145/3613904.3642116.

Leslie, David. 2019. "Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector." *Zenodo* (June 11, 2019). https://doi.org/10.5281/ZENODO.3240529.

Mehrotra, Rishabh, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. "Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems." In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management,* 2243–51. CIKM '18. Torino, Italy: Association for Computing Machinery, October 18, 2018. https://doi.org/10.1145/3269206.3272027.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts." In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,* 735–46. March 1, 2021. https://doi.org/10.1145/3442188.3445935.

Microsoft. 2022. *Microsoft Responsible AI Impact Assessment Template,* June 5, 2022. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf.

———. 2024. *Responsible AI Transparency Report,* May. https://www.microsoft.com/en-us/corporate-responsibility/responsible-ai-transparency-report.

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

Mökander, Jakob, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. 2021. "Ethics-based Auditing of Automated Decision-making Systems: Nature, Scope, and Limitations." *Science and Engineering Ethics* 27, no. 4 (July 6, 2021): 44. https://doi.org/10.1007/s11948-021-00319-4.

Moss, E., E. Watkins, R. Singh, M. C. Elish, and J. Metcalf. 2021. "Assembling Accountability: Algorithmic Impact Assessment for the Public Interest." *SSRN Electronic Journal* (June 29, 2021). https://doi.org/10.2139/ssrn.3877437.

Noble, S. U. 2020. *Algorithms of Oppression.* New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001.

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown. ISBN: 9780553418835. https://www.penguinrandomhouse.com/books/241363/weapons-of-math-destruction-by-cathy-oneil/.

OpenAI. 2023. *GPT-4 System Card.* Technical report. OpenAI, March 23, 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Perez, Caroline Criado. 2019. *Invisible Women: Data Bias in a World Designed for Men.* Abrams. ISBN: 9781419729072. https://carolinecriadoperez.com/book/invisible-women/.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 33–44. FAT* '20. Barcelona, Spain: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency.* Research report. https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2.

Rubel, Alan, Clinton Castro, and Adam Pham. 2021. *Algorithms and Autonomy: The Ethics of Automated Decision Systems.* Cambridge University Press. https://doi.org/10.1017/9781108895057.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." https://www.kevinhamilton.org/share/papers/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

Selbst, Andrew D. 2021. "An Institutional View of Algorithmic Impact Assessments." *Harvard Journal of Law & Technology* 35 (June 24, 2021): 117. https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf.

Shelby, R., S. Rismani, K. Henne, Aj. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla-Akbari, et al. 2023. "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction." In *Proceedings of the 2023 AAAI/ACM Conference on AI,* 723–41. AI, Ethics / Society, August 29, 2023. https://doi.org/10.1145/3600211.3604673.

Smith, Jessie J., Lex Beattie, and Henriette Cramer. 2023. "Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective." In *WWW '23: Proceedings of the ACM Web Conference 2023,* 3648–59. Association for Computing Machinery, April. https://doi.org/10.1145/3543507.3583204.

Stahl, Bernd Carsten, Josephina Antoniou, Nitika Bhalla, Laurence Brooks, Philip Jansen, Blerta Lindqvist, Alexey Kirichenko, Samuel Marchal, Rowena Rodrigues, Nicole Santiago, et al. 2023. "A Systematic Review of Artificial Intelligence Impact Assessments." *Artificial Intelligence Review* 56 (11): 12799–831. https://doi.org/10.1007/s10462-023-10420-8.

Stray, J., A. Halevy, P. Assar, D. Hadfield-Menell, C. Boutilier, A. Ashar, L. Beattie, et al. 2023. "Building Human Values into Recommender Systems: An Interdisciplinary Synthesis." *ACM Transactions on Recommender Systems* (June 5, 2023). https://doi.org/10.1145/3632297.

Suleyman, Mustafa. 2023. *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma.* Crown. https://www.the-coming-wave.com.

Tabassi, E. 2023. *Artificial Intelligence Risk Management Framework.* Technical report. https://doi.org/10.6028/NIST.AI.100-1.

Telefónica. 2021. *Telefónica's Approach to the Responsible Use of AI.* https://www.telefonica.com/en/wp-content/uploads/sites/5/2021/08/ia-responsible-governance.pdf.

Wang, W., and K. Siau. 2019. "Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work, and Future of Humanity: A Review and Research Agenda." *Journal of Database Management (JDM)* 30 (1): 61–79. https://doi.org/10.4018/JDM.2019010104.

Watkins, Elizabeth Anne, Emanuel Moss, Jacob Metcalf, Ranjit Singh, and Madeleine Clare Elish. 2021. "Governing Algorithmic Systems with Impact Assessments: Six Observations." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,* 1010–22. July 30, 2021. https://doi.org/10.1145/3461702.3462580.

Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. *Ethical and Social Risks of Harm from Language Models,* December 8, 2021. arXiv: 2112.04359 [cs.CL].

Wilson, Christo, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. "Building and Auditing Fair Algorithms: A Case Study in Candidate Screening." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 666–77. FAccT '21. Virtual Event, Canada: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445928.

WSJ Staff. 2021. "Inside TikTok's Algorithm: A WSJ Video Investigation." *Wall Street Journal* (July 21, 2021). https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477.

Authors

Amar Ashar (amara@spotify.com) is a Senior Researcher on Spotify's Trust & Safety team, an affiliate at the Berkman Klein Center for Internet & Society at Harvard University, and an advisor to the Trust & Safety Foundation.

Karim Ginena is the founder of RAI Audit, an AI governance and research consultancy, and Co-Vice Chair of the IEEE's AI policy committee, and he led AI fairness user research on Meta's Responsible AI team.

Maria Cipollone is an experienced user experience (UX) researcher focused on improving equity and access to quantitative analysis in UX, Data Science, and Machine Learning.

Renata Barreto is a Research Scientist at Spotify on the Trust & Safety Insights team and an affiliate at UC Berkeley's D-Lab.

Henriette Cramer is a founder of PaperMoon.AI, an AI quality and safety startup based in San Francisco. She was formerly Director of Algorithmic Impact & Responsibility at Spotify.

Acknowledgements

We are grateful to colleagues on Spotify's Trust & Safety and Product teams for their feedback, support, and collaboration. We also thank University of Washington professor Tanu Mitra for her initial review of this paper and deeply thoughtful suggestions.

Data availability statement

Algorithmic Impact Assessments were conducted on internal systems on the platform, and due to the sensitive nature of what technical and design information was collected about those systems, replication of this study externally would present trade secret and data privacy challenges. The platform includes resources on its website to help users and researchers understand how recommendations are made and what safety and responsibility measures are currently in place, including a description of its Algorithmic Impact Assessment process.

Funding statement

We thank the organization that is the focus of this study, which supported our research.

Ethical standards

User interviews took place with the explicit consent of employees within the organization described in the paper. Interview protocols were established and internally reviewed by researchers before taking place. Interview conclusions and summaries have been anonymized in order to preserve the privacy of participants.

Keywords

Algorithmic impact assessments; auditing; algorithmic impact; AI responsibility; recommendation systems; Responsible AI; algorithms; management.

Appendices

Appendix A: Summary of practitioner challenges and needs with AIAs

Practitioner Challenges	Technical and Methods	- Fairness evaluation - Harms frameworks application - Metrics, baselines, thresholds
	Infrastructure and Operations	- Legacy systems development/maintenance - Data availability - System ownership and interdependence
	Organization and Planning	- Resource prioritization - Competing commitments - Policy development and frameworks
Practitioner Needs	Applied Guidance and Governance	- Training opportunities to understand algorithmic harms - Case studies, playbooks - Golden paths - Governance frameworks for prioritization
	Internal/External Engagement	- Product embeds for applied work - Expert central teams with consultation services - Audits with second and third parties
	Operational Mitigations	- Evaluation methods, thresholds, and monitored dashboards - Guidance for internal vs. user-facing systems - Potential use of simulations to anticipate user impact