NEW AMERICA

September 2021

Cracking Open the Black Box

Promoting Fairness, Accountability, and Transparency Around High-Risk AI

Spandana Singh & Leila Doty

Last edited on September 08, 2021 at 9:18 a.m. EDT

Acknowledgments

We appreciate the many stakeholders across civil society and industry who have taken the time to talk to us about fairness, accountability, and transparency around algorithmic systems over the past year. We would also like to thank Craig Newmark Philanthropies for generously supporting our work in this area. The views expressed in this report are those of its authors and do not necessarily represent the views of Craig Newmark Philanthropies, its officers, or its employees.

About the Author(s)

Spandana Singh is a policy analyst with New America's Open Technology Institute.

Leila Doty was a Legal/Public Policy intern with New America's Open Technology Institute, working with the platform accountability and secure teams.

About New America

We are dedicated to renewing the promise of America by continuing the quest to realize our nation's highest ideals, honestly confronting the challenges caused by rapid technological and social change, and seizing the opportunities those changes create.

About Open Technology Institute

OTI works at the intersection of technology and policy to ensure that every community has equitable access to digital technology and its benefits. We promote universal access to communications technologies that are both open and secure, using a multidisciplinary approach that brings together advocates, researchers, organizers, and innovators.

Contents

Contents Cont'd

Introduction

This report will provide a landscape overview of some of the most prominent and promising proposals related to how internet platforms and governments can create and institute mechanisms for promoting fairness, accountability, and transparency (FAT) around high-risk algorithmic systems, with the goal of discussing the strengths and limitations of these approaches and demonstrating how these mechanisms fit into the overall FAT landscape.

Over the past decade, private companies and government agencies have radically expanded their development and use of machine learning (ML) and artificial intelligence (AI). While these algorithmic systems can allow entities to operate with greater efficiency and scale, they can also generate discriminatory, biased, and otherwise harmful outcomes. These harms can have wide-ranging consequences and have occurred across industries, including in education,[1] labor,[2] medicine, and criminal justice sectors,[3] and throughout online platforms.[4] Federal privacy legislation—which is direly needed for a variety of reasons—would likely help to address many of the existing harms stemming from private sector use of algorithmic systems today, but may still need to be supplemented with additional policy measures focused on algorithms and algorithmic accountability.[5]

In response, civil society organizations and civil rights groups, researchers, and policymakers have begun to think about how to promote greater FAT around the use of algorithmic systems, especially systems that pose high risks to citizens and society. These conversations have thus far resulted in the development of numerous high-level principles and guidelines around ethical uses of AI. While helpful for identifying critical values to consider, these types of outputs are limited by the fact that they are hard to translate into practice.[6] In addition, many discussions around how to deploy certain FAT mechanisms occur in a siloed manner, failing to account for the fact that numerous mechanisms must often be deployed in concert to effectively promote meaningful FAT around these systems throughout different parts of their life cycle.

In addition, numerous existing considerations of FAT fail to account for whom a FAT mechanism is designed. In order for a FAT measure to be effective, it must be meaningful and explainable. If the information being disclosed is not comprehensible by the intended end user, then it is not a valuable mechanism for promoting FAT.[7] Further, thus far there has been little consensus around how to define a high-risk algorithmic system; it would be helpful for researchers, civil society organizations, and other key stakeholders to provide an agreed-upon definition to companies and government entities. This will also ensure that any assignment of responsibility or liability is proportionate.

As this report will outline, there are a wide variety of mechanisms for promoting FAT around high-risk algorithmic systems that seek to counteract the harmful effects of opacity and address concerns around bias and discrimination. This report will unpack nine different categories of approaches, including their strengths and weaknesses, and outline best practices for using some of these mechanisms. These nine categories were selected based on their prominence in ongoing conversations around promoting FAT around high-risk algorithmic systems.

This report contains four sections, each outlining approaches to promoting FAT around high-risk algorithmic systems that different entities would be best suited to implement or pursue. Section one discusses internet platforms, section two discusses government entities and regulators, section three discusses internet platforms and government entities, and section four discusses other stakeholders.The report concludes with recommendations on next steps that internet platforms and governments deploying and using algorithmic systems should prioritize, as well as recommendations for future multi-stakeholder engagement. The recommendations are similarly broken down by which actors are best suited to implement them. Throughout the report, we reference research and examples from the European Union (EU) and other regions to inform our analysis. However, our recommendations are primarily focused in the U.S. context. This report builds on OTI's report and event series—*Holding Platforms Accountable: Online Speech in the Age of Algorithms*—which explores how internet platforms use algorithmic decision-making to shape and curate the content we see and makes recommendations on how platforms can promote greater FAT.[8]

Editorial disclosure: This brief discusses policies by Facebook, Google (including YouTube), Microsoft, and Twitter all of which are funders of work at New America but did not contribute funds directly to the research or writing of this piece. View our full list of donors at www.newamerica.org/our-funding.

FAT Approaches Internet Platforms Can Implement

Amid the various ways to promote FAT around high-risk algorithmic systems, there are two that internet platforms are best suited to implement: ML documentation frameworks and transparency reports. This section includes a discussion of three different types of ML documentation frameworks (Datasheets for Datasets, Model Cards, and FactSheets), and two types of transparency reports (content takedown transparency reports and political ad transparency reports). It discusses the strengths and limitations of these two approaches, and how these mechanisms contribute to overall efforts to promote FAT around high-risk algorithmic systems.

Machine Learning Documentation Frameworks

Currently, numerous industries implement standardized documentation methods to communicate the function and quality of a given system. For example, civil engineers use standardized engineering drawing practices to prepare structural plans that can be understood across the industry.[9] However, despite some proposals (discussed below), there are currently no standardized documentation procedures in the ML community,[10] hindering efforts to promote FAT around ML and algorithmic systems. As noted previously, documentation throughout the ML training lifecycle is important because it allows developers to track, revisit, and understand past design decisions and enables external reviewers to conduct a substantive audit of the algorithmic system. Although internet platforms have been the most prominent target for such a documentation framework, these methods could also be extended to other commercial and even government uses.

Datasheets for Datasets

In 2018, a team of Microsoft researchers proposed Datasheets for Datasets, a documentation framework primarily for companies in which each dataset in the ML process is accompanied by a datasheet that expresses the dataset's motivation, composition, collection process, and recommended uses, among other characteristics.[11] The Datasheets for Datasets framework helps facilitate meaningful communication between the creators and the consumers of datasets, encouraging creators to carefully reflect on the process of creating, distributing, and maintaining a dataset and empowering consumers with the facts they need about it. This is important given that flawed datasets can generate harm within an ML model or algorithmic system. However, should an individual use the dataset outside of the creator's intended use and guidance, there is no clear mechanism for recourse.

FactSheets

Researchers at IBM have proposed the FactSheets framework, which relies on documents called FactSheets to collect relevant information about the development and deployment of an AI model or service in a common, transparent location. The documented information includes the purpose of the model, the dataset used to train the model, limitations of the model's performance, and many other factors. Throughout the life cycle of the AI model, facts are recorded by various stakeholders in the process, including the business owner, data scientist, model validator, and operations engineer. FactSheets are tailored to the particular documentation needs of different audiences (e.g., model developers, regulators, consumers, etc.) and therefore may vary in content and format (e.g., full report, tabular view, slide format, etc.) depending on the targeted audience, even for the same AI tool.[12]

According to IBM, FactSheets enable the governance of algorithmic systems by providing enterprises the ability to track and understand information throughout the AI life cycle, analyze this information, specify policies to be adhered to during AI development and deployment, and facilitate communication and collaboration among stakeholders at various points in the AI life cycle.

Model Cards

Finally, researchers at Google have proposed using "model cards" to clarify the intended use cases of ML models and curtail the use of these models in situations for which they are not appropriate. Model cards are short documents which accompany trained ML models, providing evaluations of the models that are related to a number of conditions, including how the model fairs across different cultural, demographic, or relevant phenotypic groups. Model cards also outline the intended use cases of the model and include information on the performance evaluation procedures,[13] the motivation behind chosen performance metrics, group definitions, and other factors that relate to bias, fairness, and inclusion.

Model cards can provide value to a range of stakeholders: they can inform policymakers on when an ML system would operate appropriately and when the system would fail, help permit developers to compare and contrast model results to better train their own systems, and empower individuals who have been negatively impacted by a system to understand how it works and how to pursue remediation.[14] Model cards can thus serve as a valuable mechanism for promoting transparency between developers, users, and other stakeholders around how automated systems are developed and deployed across industries. The framework is also flexible and can therefore be applied in different contexts, to different stakeholders.[15] Thus far, model cards have primarily been explored in the corporate context by internet platforms and associated researchers.[16]

However, model cards are limited in that their value is dependent on the creator of cards. If the card creator is not fully transparent, the model card will not be useful. Further, although a model card creator may outline their intended use

cases for a model, there is nothing stopping another individual or entity from deploying a model in inappropriate ways.[17]

Datasheets for datasets, fact sheets, and model cards are important steps toward universal documentation practices and promoting FAT around algorithmic systems—some combination of them may be useful—but individually there is still room for improvement in each framework. In comparison to fact sheets, model cards are missing some critical information about an ML model, such as robustness (how stable the model's performance is in a variety of environments) and substantive information about bias present in the model.

The Partnership on AI (PAI), a multi-stakeholder organization that aims to identify best practices for AI development and deployment, is working on increasing transparency and accountability around internet platforms' use of algorithmic systems with ML system documentation through their ongoing initiative called the Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles (ABOUT ML).[18] ABOUT ML seeks to synthesize best practices for data and systems documentation, building on datasheets for datasets, fact sheets, and model cards. These kinds of multi stakeholder efforts are promising, but they need to be inclusive and account for stakeholder power dynamics in order to generate meaningful outcomes.

Transparency Reports

Over the past decade, transparency reports have become an effective mechanism for obtaining transparency and accountability from internet platforms and companies in other industries. Originally, internet platforms published transparency reports to outline the scope and scale of government requests for user data they received.[19] This practice became more common after the Snowden revelations in 2013, after which a broad range of internet and telecommunications companies began issuing reports. Today, transparency reporting on government requests for user data is considered an industry-wide best practice for internet platforms.

Over the past several years, internet platforms and some telecommunications companies have expanded their transparency reporting to include data on government requests for content takedowns and network shutdowns.[20] More recently, a handful of internet companies that host user-generated content have also begun reporting on how they enforce their content policies when moderating content.[21]

Some of these transparency reports include data that relates to how these companies use AI, but only in the content moderation context. For example, Facebook's *Community Standards Enforcement Report* (CSER) includes a metric called "proactive rate," which outlines how much content the platform

proactively removed using human reviewers and automated tools (rather than relying on user reports).[22] However, the metrics shared are relatively broad, and therefore provide limited insight into how the platform uses AI and ML for content detection and moderation. Similarly, YouTube details the number of videos and comments it has removed and breaks these figures down by source of first detection in its community guidelines enforcement report.[23]

Although internet platforms rely on AI and ML tools to facilitate their content moderation processes, these companies provide very little transparency around how these tools are used and what impact they have.[24] Additionally, out of the handful of companies that issue transparency reports on how they enforce their content policies, only Facebook and Twitter report on their use of automated tools.[25]

After the 2016 U.S. presidential election, internet platforms began publishing a new type of transparency report that covered their online advertising operations. This new form of reporting emerged in response to calls for platforms to be more transparent around their algorithmic advertising systems, especially after these advertising systems were used to sow discord during the electoral cycle.[26] In addition, platforms such as Facebook have received additional scrutiny from policymakers and researchers after investigations revealed its algorithmic ad targeting and delivery service enabled the discriminatory targeting of ads, including in the housing, employment, and credit contexts.[27]

In response to mounting pressure from policymakers and civil society on platforms to provide transparency around their algorithmic advertising operations, companies such as Facebook, Google, and Reddit started issuing ad transparency reports, also known as ad libraries. Facebook's ad library includes information on its housing, employment, credit, issues, elections, and political ads. Google and Reddit's reports focus on political ads.

Although these ad transparency reports are a valuable first step toward providing insight into how these companies' ad targeting and delivery systems work, they offer very little granular information on the algorithms themselves. Facebook's ad library includes information about the impressions an ad has accrued (Facebook defines an impression as the number of times an ad was seen on a screen) and aggregate information on the age and gender of users who were shown an ad, among other metrics.[28] Google's political-ad transparency report includes data on impressions and the targeting criteria an advertiser selected before running an ad.[29] Reddit's subreddit r/RedditPoliticalAds offers similar data on impressions and geo- and subreddit targeting.[30] None of these reports, however, provide granular information on the reach and engagement of an ad (e.g., how many likes, shares, and video views an ad received) or a comparison of what audience segments the ad initially targeted and what audience segments eventually received the ad.[31] This information is critical to understanding the role advertising algorithms play in facilitating online harms, including discrimination.

Companies—including Facebook—have stated that they are unable to share such granular information publicly due to privacy concerns.[32] However, numerous researchers have suggested safeguards that could enable these disclosures to be made responsibly.[33]

Thus far, transparency reports have been a valuable mechanism for obtaining aggregate data from internet platforms around their privacy and freedom of expression commitments, but improvements are needed for this practice to be a valuable method for promoting FAT around platform use of algorithms. Companies need to report metrics that are more directly related to the use, accuracy, and impact of algorithmic tools in content moderation and digital advertising systems, such as error rates. One of the challenges here is that platform use of automated tools is not always binary—when moderating content, a platform may use automated tools to detect content and route it to a human reviewer who will make the relevant decision, or vice versa—so crafting meaningful metrics that capture when automated tools are in use and when they are not can be difficult. However, companies can publish qualitative information explaining how and when they use these tools during the content moderation process nonetheless. Platforms can also publish quantitative data that outlines when automated tools were used for certain purposes, such as to flag content or to remove it. In addition, transparency reporting is an expensive and labor-intensive process, especially considering that not all platforms have the necessary data collection structures already in place. As a result, some platforms have asserted that they have to triage requests for new metrics they receive and prioritize ones they determine will provide the most meaningful transparency (and are beneficial to their own business priorities).

However, these challenges can be overcome to develop transparency reporting into an effective mechanism for promoting FAT around both high-risk systems and lower risk AI systems. Recognizing these obstacles, civil society and advocates should come together to identify a set of metrics related to content moderation and digital advertising that platforms should prioritize when issuing transparency reports. One example of this is the *Santa Clara Principles on Transparency and Accountability in Content Moderation*, which include some metrics that touch on the use of automated tools during content moderation.[34] Ranking Digital Rights's *Corporate Accountability Index* also includes numerous indicators related to transparency reporting and the use of automated tools for content moderation and curation.[35] As advocates continue to think about key metrics in this regard and in relation to other forms of content curation, such as content ranking and recommendation systems, they should identify a set of priority metrics that platforms should initially focus on and how they can expand on those in the future.[36] Such conversations will also help clarify what aggregate data would be most meaningful to have in order to promote algorithmic FAT around the use of content moderation and digital advertising systems.

Policymakers in the European Union[37] and the United States[38] are also exploring whether transparency reporting around content moderation and digital advertising should be mandated by legislation, and these conversations should continue and be given proper attention. As noted above, some experts have also suggested establishing a regulator that would focus on transparency disclosures and information sharing around algorithmic systems that have been deemed as warranting additional scrutiny.[39] Transparency reporting could be part of these proposed regulatory efforts.

Internet platforms should also supplement transparency reporting efforts by publishing accessible and easy to understand algorithmic-system use policies, which explain to consumers how the company uses algorithms and for what purposes. Companies should also enable users to determine whether and how their personal data is used to train these systems, to understand what data points are used to inform the companies' algorithmic systems, and to opt out of using these systems altogether where possible (e.g., they should be able to opt out of receiving an algorithmically curated news feed). Further, companies should establish mechanisms for consumers to provide feedback on adverse outcomes that result from an algorithmic system. Some platforms currently offer this for content moderation decisions, which can be made using algorithmic systems. On platforms such as Facebook, Twitter, and YouTube, users can appeal the suspension or removal of their content and accounts in certain cases.[40] This is also an important accountability mechanism.

FAT Approaches Governments Can Implement

There are two approaches to promoting FAT around high-risk algorithmic systems that governments are best suited to implement: enforcement mechanisms and procurement guidelines. This section provides an overview of these two mechanisms and discusses their strengths and limitations, and how these approaches contribute to overall efforts to promote FAT around high-risk algorithmic systems. This section includes a discussion of several different enforcement mechanisms governments can take in order to promote greater FAT around high-risk algorithmic systems. These include establishing a regulatory body or agency dedicated to algorithmic accountability issues, taking executive action, and passing legislation. This section also outlines how governments can utilize the public procurement process to incentivize the development of algorithmic systems that reflect adequate FAT.

Enforcement Mechanisms

Enforcement is a critical component of the FAT process, as it provides accountability and can promote transparency. Policymakers across the globe, including certain U.S. legislators and governments in the EU, have been exploring regulatory and legislative action around algorithmic systems. As governments begin exploring how to hold relevant actors accountable for harms caused by AI systems, it is important that they consider the geographic scope (and limitations) of any proposed efforts, and that algorithmic systems are likely deployed in a multinational manner, beyond where one country can enforce regulations on the back end. This is critical for ensuring that any legislative methods for promoting FAT are effective.[41]

In addition, should governments choose to pursue a regulatory model for high-risk algorithms that are deployed by internet platforms and government agencies, they must ensure that any regulatory action considers the different components and actors of the AI-system life cycle, operates using a clear, consensus-based definition of high-risk algorithmic systems, and assigns responsibility and liability for damages to the actors that are best suited to address potential risks.[42] Depending on the algorithmic system, regulatory action could target the developer, the entity or individual deploying the product, or other actors, such as the distributors or importers, service providers, or end users.[43]

Some methods of enforcement and regulation that governments, researchers, and civil society organizations have proposed include:

1. Forming a regulatory body that imposes binding regulation on entities that develop or deploy high-risk algorithmic systems. Experts argue that because algorithmic systems are opaque and can generate small but severe long-term harms across industries, they should be subject to additional regulatory scrutiny.[44] Such a regulator could require that private companies and government agencies obtain pre-market approval from the agency before deploying an algorithmic system and could conditionally approve the use of an algorithm in limited circumstances. If an entity deploys an algorithm outside of the agreed upon circumstances, the regulator could subject the entity to legal consequences.[45] The regulator could also allow injured parties to obtain damages.[46]

   While this approach could provide accountability in situations where high-risk algorithmic systems generate significant harm, it is limited by the fact that algorithmic systems are continuously evolving based on new data and inputs.[47] As a result, it is difficult to predict what exact outputs an algorithm would produce in a given situation and how to cleanly draw lines between high-risk algorithmic systems and systems that pose less of a threat.[48]

   Alternatively, the European Parliamentary Research Service has suggested that a central regulatory body for algorithms be empowered to focus on three critical characteristics of algorithmic systems: complexity, opacity, and dangerousness. Experts argue that a central regulatory agency (sometimes discussed as an "FDA for algorithms")[49] would be better equipped than individual agencies to tackle incidents of algorithmic harm, as it would be able to centralize talent and develop more comprehensive best practices and review procedures.[50] Although establishing a regulatory body could help promote FAT around the development and deployment of algorithmic systems, the process of establishing such a body would be lengthy, require significant investment of resources and talent, and would not address concerns around FAT in the short term.

2. Creating an agency responsible for certifying the safety of an algorithmic system. Under this proposal, the certifying agency would rely on a legal liability framework that subjects developers and vendors of these certified systems to tort liability.[51] Algorithmic systems that are uncertified and commercially sold or used would be subject to stricter liability. Accordingly, courts would be responsible for deciding whether an algorithmic system is within the scope of the agency certification process and for assigning responsibility to relevant actors when a system produces tortious harm. As some experts have noted, this type of regime would encourage developers to think more critically about the costs associated

with algorithmic system harms. It would also enable victims of any harm to seek compensation.[52]

3. Relying on consumer protection authorities, such as the Federal Trade Commission (FTC) to apply consumer protection regulations to user agreements, therefore generating greater accountability for operators.[53] In addition, the FTC could push developers and deploying entities to provide greater transparency around their algorithmic systems, which could include algorithmic audits, to facilitate the FTC's own evaluations. The FTC could also help generate accountability by requiring regular reporting to the agency. Reporting that is not required publicly could help address company claims that providing too much transparency around their algorithmic systems could amount to giving away trade secrets while also enabling oversight and mitigation of potential harms caused by these algorithmic systems.[54] This level of transparency, however, does not increase public insight into how these systems work and what impacts they may generate.

4. Establishing a government-led incentive or penalty-based system in which entities provide subsidies or funding to adopt certain FAT practices or are taxed, fined, or made to pay fees if they do not. In other industries, governments have created tax incentives to encourage responsible corporate behavior,[55] such as those used to promote the use of environmentally friendly technologies, including electric cars and solar energy.[56] A similar structure could be used to encourage internet platforms to implement FAT measures including voluntary labeling or certification schemes, and algorithmic audits. This approach could be useful for addressing lower-risk algorithmic systems that don't particularly require mandatory guidelines around how they are used and what FAT mechanisms need to be implemented in order to offset their risks.[57] However, given the vast potential harms that could arise from internet platform development and deployment of high-risk algorithmic systems, companies should pursue measures to promote FAT independently regardless. For example, the use of taxes, fines, and fees would make the most sense if applied to high-risk algorithmic systems. However, the policies that determine if and when an entity is fined must be detailed and clear, and must carefully account for potential impact on smaller companies and therefore market competition. Any efforts to impose penalties on companies operating algorithmic systems must be careful to not produce any unintended harms to fundamental rights.[58]

In April 2021, the European Commission released a draft of its proposed AI regulation, which seeks to rein in and prohibit certain uses of high-risk algorithmic systems.[59] While the proposal makes some notable strides, its

provisions do not clearly and broadly implicate internet platform use of these systems. In comparison, the EU's draft of the Digital Services Act (DSA),[60] released in December 2020, better addresses internet platforms' use of algorithmic systems that could generate harms, including targeted advertising and recommendation systems. The most recent draft of the DSA includes provisions that would require internet platforms to provide greater transparency and user control around such algorithmic systems.[61]

The European Commission's draft AI regulation also includes provisions that impact government use of high-risk algorithmic systems. For example, it introduces prohibitions on certain uses of AI-based social scoring systems by public authorities and on the use of "real-time" remote biometric identification systems by law enforcement in public spaces.[62]

The U.S. government has also taken some high-level steps to promote FAT around algorithmic systems. In 2019, the government introduced an Executive Order (EO) that aims to establish and maintain U.S. leadership in AI.[63] While the EO does not offer granular guidance on mechanisms for promoting FAT it does encourage the government to "train current and future generations of American workers with the skills to develop and apply AI technologies."[64] One of the barriers to establishing processes and implementing mechanisms for promoting FAT around algorithmic systems at the government level is that government agencies often lack the necessary technical talent. The EO provides valuable high-level guidance that the government should institute mechanisms and programs for recruiting and upskilling technical talent and staff in government agencies. In addition, in response to the EO, the National Institute for Standards and Technology (NIST) released a plan calling for the federal government to engage in long term AI standards development activities. NIST has also stated that it will collaborate with the private sector and academia to establish AI standards which address broader societal, governance, and privacy issues.[65] These standards are still in development, but they could offer valuable guidance for internet platforms and government agencies as they seek to promote FAT around their development and use of algorithmic systems in the long term.

In 2020, the U.S. government issued another EO, which aims to promote the use of trustworthy AI in the federal government.[66] This second EO establishes a set of high-level principles to which federal government use of AI must adhere, including reliability, safety, security, resiliency, transparency, and accountability.[67]

Going forward, the U.S. government should promote greater FAT by issuing an EO that requires both internet platforms and government agencies to evaluate any high-risk algorithmic system before it is deployed. The EO should require these systems be subject to continuous and periodic reviews to account for changes in systems, how they operate, and what risks they produce. In addition, the U.S. government should supplement this EO with clear rules that require

companies and government agencies to review their algorithmic systems—particularly their high-risk algorithmic systems—before they are deployed and to mitigate any identified harms. If these entities fail to do so, they can be held liable by a regulator.

In addition, the U.S. government should pass comprehensive federal privacy legislation that includes specific references to the development and use of algorithmic systems. Such privacy legislation should also require transparency, impact assessments, and regular audits from internet platforms to prevent algorithmic tools from being used in ways that disparately impact disadvantaged communities.[68] These rules should also empower the FTC or a new Data Protection Authority to enforce requirements and develop regulations. As previously noted, federal privacy legislation may still need to be supplemented with additional policy measures, such as a standalone bill that requires FAT mechanisms like algorithmic audits and impact assessments to prevent abusive users of algorithmic tools and mitigate discriminatory harms.[69]

Procurement

The public procurement process presents an opportunity for governments at all levels—local, state, and federal—to incentivize the development of AI-enabled technologies and services that are fair, accountable, and transparent.[70] In fiscal year 2019 alone, the U.S. government spent over $20.7 billion on information technologies, computer software, and engineering-related services, including AI-powered technologies.[71] Because it can be difficult for the government to develop algorithmic systems internally—due to high costs and not enough skilled in-house technologists—government agencies often acquire AI-enabled tools through public procurement processes. According to a recent report commissioned by the Administrative Conference of the United States (ACUS), 47 percent of AI in use among the federal government was developed externally.

On a basic level, the public procurement process consists of a government agency identifying the need for a good or service, issuing a request for proposal (RFP), seeking responses from companies until a closing date, and entering into a contract with the lowest bidder.[72] In the United States, companies that seek to win an agency's contract must meet the basic quality standards required by law, in addition to the context-specific safety and performance requirements indicated in the RFP. Unfortunately, current procurement standards are outdated and insufficient for regulating emerging technologies, but policymakers could update these standards and outline clear requirements for promoting FAT around algorithmic systems within them.[73] By requiring FAT-promoting mechanisms such as audits of algorithmic models, algorithmic impact assessments, and disclosure of training and testing data in an RFP, the public procurement process could effectively promote FAT around the government's algorithmic systems.

The European Commission's High Level Expert Group on Artificial Intelligence recommended the strategic use of public procurement to fund innovation and develop trustworthy AI by ensuring that governments identify, assess, and appropriately address potential risks, and add eligibility and selection criteria for algorithmic systems.[74] In addition, the City of Amsterdam has drafted "standard clauses for municipalities for fair use of algorithmic systems" that seek to operationalize ethical AI principles intended to be included in the procurement contract of any government acquisition involving AI technologies.[75]

Although procurement processes present a powerful demand-side opportunity for the government to incentivize the private development of FAT around AI, this mechanism also has its limitations. First, public procurement only directly impacts the government's use of algorithmic systems and does not guarantee that private sector development and deployment of algorithmic systems outside of a government context will also promote more FAT around AI. Second, for "soft/custom" goods and services that undergo the procurement process, like tailored AI tools (e.g., an algorithm that attempts to assign students in a city to schools in such a way to make each school's population geographically and racially diverse), achieving a high-quality product involves specialized skills and adequate knowledge of the deployment context. Contracted engineers from the private sector may not possess a nuanced understanding of the problems an algorithmic system is aimed to address, including the complex legal, regulatory, and organizational environment in which the tool will be deployed. Where agencies must contract out the development of an AI tool, in-house agency personnel who have expertise on the problem they seek to address should work closely with the contracted private sector experts to provide the necessary contextual understanding.

Overall, public procurement is a critical tool to promote FAT around AI use. However, public procurement only directly impacts public use of algorithmic systems so the development of these systems should ideally happen in-house to allow for greater quality control and auditing. That said, as long as the government contracts with private vendors, procurement standards should be updated to regulate emerging technologies as much as possible. Further, as policymakers consider the role government agencies can play in overseeing the use of algorithmic systems, it is critical that these entities increase their in-house technical expertise so they can adequately carry out these functions.[76]

Internet platforms also have the ability to procure AI tools and services, but are largely free to formulate their own bidding processes, which usually results in much less transparency in private procurement practices.[77] Since the private procurement process is much less regulated than that of public procurement, it is a much less effective opportunity to promote FAT around algorithmic systems through procurement practices at scale.

FAT Approaches Internet Platforms and Governments Can Implement

This section discusses four approaches to promoting FAT around high-risk algorithmic systems that both internet platforms and governments can implement: Algorithmic Impact Assessments, algorithmic audits, bias impact statements, and labels for algorithmic systems. This section explores the strengths and limitations of each of these approaches, as well as how these mechanisms individually fit into the broader ecosystem of approaches seeking to promote FAT around high-risk algorithmic systems.

Algorithmic Audits

Despite the ubiquitous use of algorithmic systems across industries today, both experts and the general public still have a limited understanding of how these tools work. This is because algorithms operate in an opaque manner. Depending on the algorithm, developers can know what inputs are fed into the system and what outputs are generated. However, even developers themselves often have very little insight into the inner workings of the algorithm. In this way, algorithms can function as black boxes.

Algorithmic audits can help address the opaque nature of algorithmic systems by allowing auditors to evaluate and scrutinize the inner workings of an algorithmic system as it is being deployed or after it has been deployed. Audits can help evaluate specific variables, such as those related to privacy, human rights impacts (e.g., freedom of expression), bias, and fairness. Audits can also help identify unintended consequences,[78] examine concerns raised by external stakeholders (e.g., civil rights groups), and/or determine whether a system is aligned with certain company or government policies, industry standards, or regulations.[79] Depending on how an audit is conducted, how transparently the results are published, and whether the audited entity takes meaningful steps to mitigate any problems identified during the audit, audits can be an effective tool for promoting greater FAT around algorithmic systems that are being implemented or are already in use.

Algorithmic audits can be useful in both a corporate and government context. The concept of auditing companies is a longstanding one that has transformed over the decades.[80] Today, auditing is a critical mechanism for quality control in sectors such as aerospace and healthcare. It is also a common practice for private companies to submit to financial audits.[81] The practice of auditing for discrimination has also been in place in the public sector since the 1970's, when the research unit of Department of Housing and Urban Development (HUD) carried out audits to detect racial discrimination in housing.[82] As more

companies and government entities rely on algorithms to share information about and make key decisions related to housing, finance, and other consequential areas of life, experts have proposed auditing algorithmic systems as a mechanism for promoting accountability around these tools.

Currently, the biggest issue with all algorithmic audits is there is no established algorithmic auditing structure or landscape. In the financial sector, there are well-established auditing standards. If an auditing company or entity being audited seeks to evade transparency and accountability, both the auditor and entity being audited will face reputational damage and legal liability. However, in the emerging algorithmic auditing space, no clear norms have been established, so no harms exist for either party. This vagueness could serve as a disincentive for companies and government entities to participate in algorithmic audits at the current moment.[83] This is an area where thoroughly researched standards or government oversight can help promote greater FAT via algorithmic audits.

Audits on a company or government agency's algorithms can either be conducted internally (i.e., by the company or agency themselves) or externally (i.e., by an independent third party). Researchers have noted that internal audits can be beneficial mechanisms for promoting FAT around algorithmic systems, as auditors will have access to a robust set of information on the relevant system. Internal auditors will also likely be better versed in the entity's operations and technical infrastructure compared to external auditors. In order to promote transparency and reliability, companies must address any issues identified during internal audits and should publish at least an explainable summary of the audit's findings to the public. If performed legitimately, internal audits could supplement external transparency and accountability efforts, including external audits.[84] In other industries, such as the financial, chemical, food, and aviation sectors, for example, internal auditing practices related to quality assurance are also coupled with regulatory mechanisms, which guide expectations and standards around internal audits.[85] Without external checks like these, it is unlikely that company or government audits of their own internal systems will accrue legitimacy, as they allow the entity to essentially create their own tests and grade their own homework.

In the same vein, external audits conducted by independent third parties are likely to be more reliable and legitimate, but are limited by the current lack of algorithmic auditing standards, which will hamper external parties' access to and understanding of an entity's internal processes. When it comes to private companies, external auditors may only be able to access model outputs through alternative avenues, such as application programming interfaces (APIs). They also may not have access to critical information, such as intermediate models or training data, as these are often shielded as trade secrets by intellectual property claims. However, external auditing is widely accepted for practices such as financial auditing, and often relies on the use of non-disclosure agreements

(NDAs). As standards for algorithmic auditing are developed, private companies should adopt a similar external auditing structure.

Given the current limitations around conducting external audits on internet platforms' algorithms, some researchers have proposed alternative avenues for conducting such audits in the short term, particularly in situations where researchers may not have full access to a company's systems. In many of these instances, researchers do not have consent from companies to conduct these audits. This poses significant legal risk to researchers, limits the effectiveness of the audits, and underscores the need for auditing standards. Some of the alternative auditing methods proposed include:

1. **Code audits**, in which a platform discloses its source code to researchers or the public. However, given company concerns including trade secrets, adversarial use of their algorithms, and the privacy of their users, companies would likely only provide such disclosures where they are compelled by the government. In addition, this approach is limited in that reading source code does not immediately facilitate the interpretation of algorithms or the identification of harmful outcomes. Rather, an algorithm's outputs are reliant on its inputs. As a result, researchers would have to rely on trial and error to identify harms when given just source code. These limitations also underscore the fact that algorithmic systems *and* flawed datasets lead to harmful outcomes, and must be considered together.

2. **Noninvasive user audits**, in which users agree to answer questions about—or provide researchers access to—data on their online behaviors so that inferences can be made about the operations of an algorithm. However, this approach does not involve actually testing the algorithm in any way, and is vulnerable to sampling issues as well as high error rates common with self-reporting mechanisms for data collection.

3. **Scraping audits**, in which a researcher could query a platform and evaluate the results. This is often done through an API. However, in the United States, the Computer Fraud and Abuse Act (CFAA) creates significant legal risks for researchers even though its purpose is to criminalize hacking.[86] Many platforms also include stipulations that hinder research efforts in their terms of service.

4. **Sock puppet audits**, in which researchers rely on computer programs to impersonate users on a platform. Because this approach requires deception, researchers or those creating the programs can incur similar legal consequences under the CFAA. The operator of an algorithm could also claim that the use of sock puppet accounts is harmful as they perturb

an algorithm and could undermine its operations. Researchers therefore have to tread carefully if deploying this method.

5. **Crowdsourced/collaborative audits**, in which users volunteer or are hired to perform tasks online to test an platform's algorithms. This approach can be costly, but likely does not incur legal consequences under the CFAA.

In order to understand the role algorithmic audits can play in promoting FAT around algorithmic systems, it is important to consider their strengths and limitations. A sizable amount of an audit's legitimacy will be derived from if and how the auditing entity and the entity being audited communicates the audits results. If condensed information about the outcomes of audits are published in an explainable manner and met with oversight from a relevant body,[87] it could help boost awareness around the potential harms from algorithmic systems and engender trust that they are being mitigated. However, given the current unstructured and non-compulsory auditing environment, companies and government agencies may be reluctant to voluntarily audit their algorithms and share outcomes, as they likely fear negative reactions. But, if the results of audits are kept entirely private, and there are no methods for oversight, there is also no way to ensure that companies and governments are being held accountable.

In order for algorithmic auditing to become a reliable mechanism for promoting FAT, relevant stakeholders—such as policymakers, civil society groups, and standards-setting bodies—need to develop appropriate standards of practice, training and credentialing for auditors, transparency conventions, and other mechanisms that essentially turn this practice into a professional field.[88] The creation of standards for algorithmic auditing by relevant stakeholders is important for a number of reasons. Thus far, algorithmic auditing has been carried out by a range of actors, such as investigative journalists. However, without a set of conventions to guide how these audits are conducted, it is difficult to compare, contrast, and verify the results of audits.[89] Audits are also dependent upon human judgment, and they can therefore vary in their reliability.[90] Standards can help combat this. Additionally, in their current form, algorithmic audits often seek to address different values and concerns (e.g., discrimination, media plurality, etc.) and integrate concepts from different disciplines (e.g., human-centered design, behavioral economics, ethics, etc.[91] These concepts and values are varied and can be subjectively defined. In order to mitigate subjectivity, audits must be designed and deployed using a clear, standardized methodology and process. These standards should also guide disclosure and transparency expectations and clearly define high-risk algorithmic systems in a manner that accounts for the fact that the risks an algorithmic system poses over time can vary. Clear standards will also guide companies and governments on how to design their algorithmic systems on the back end so that they are compatible with audit mechanisms.

As policymakers or standards-setting bodies seek to develop standards for algorithmic audits, they should also consider the different types of algorithmic systems that companies and government agencies operate, their use cases, and their potential to cause harm and create high-risk situations. It is also important for these actors to recognize that algorithmic systems are not static. They are constantly being retrained and redeployed. Accordingly, any efforts to encourage reviews of algorithmic systems, such as audits, must include plans for ongoing accountability, not just one evaluation.[92]

Algorithmic Impact Assessments

Algorithmic Impact Assessments (AIAs) are another mechanism that seeks to promote FAT around algorithmic systems. AIAs evaluate algorithmic systems pre-deployment to help determine whether their use is appropriate in a given context by documenting the potential impact of the system.[93] In general, AIAs are meant to be used by government agencies or private companies that intend to deploy AI systems as self-evaluations to identify the potential harms of the system and provide a holistic view of the impacts.

One of the first pieces of legislation attempting to regulate algorithmic systems leans heavily on AIAs to promote AI accountability: the Algorithmic Accountability Act. Proposed in the U.S. Congress in 2019, this legislation would require large companies to conduct impact assessments of the automated decision systems they deploy that may affect sensitive personal information.[94] Notably, however, the bill does not provide meaningful details regarding how those AIAs should be structured or implemented. The AI Now Institute, the European Parliament, and the Canadian Government have all proposed versions of AIAs for governments and companies that draw directly from long-standing impact assessment frameworks in other policy domains, such as environmental protection, human rights, privacy, and data protection.[95]

At present, there is no consensus around what specific elements an AIA should contain. Current proposals generally suggest that AIAs should include a statement explaining the algorithmic system's potential impacts and share relevant and widely interpretable details of how the algorithmic system operates. The AI Now Institute has proposed a useful framework for government implementation of AIAs, recommending that they be incorporated into the public procurement process so that the government can provide greater accountability to the public around its use of algorithmic systems. Under their proposal, AIAs would consist of five key elements: 1) government agencies would conduct a self-assessment of their existing and proposed automated decision system, evaluating impacts on bias, fairness, and justice; 2) agencies would develop a meaningful external researcher review process; 3) agencies would disclose their definition of automated decision system to the public; 4) agencies

would solicit public comments to clarify concerns and address outstanding questions; and 5) the government would provide due process mechanisms for affected individuals and communities to challenge inadequate agency self-assessments or harmful uses that an agency fails to mitigate or correct. The AI Now proposal recommends that AIAs be incorporated into the pre-acquisition stage of procurement processes, so that the agency can evaluate the adoption of an automated decision system and take public input into account before committing to its use.[96] Other advocacy groups, including OTI, have suggested that AIA's should assess elements central to traditional privacy legislation, such as data minimization, retention periods for personal information, and whether users can access, challenge, or correct decisions made by an algorithmic system.[97]

As some researchers have laid out, AIAs could be valuable for promoting FAT around algorithmic systems that the government or a company seeks to use. As AI Now suggests, placing AIAs at the pre-acquisition stage of procurement would inform the public of the automated decision system's functions and potential impacts, allowing them to identify concerns that may need to be negotiated or otherwise addressed *before* a contract is signed. This could allow the government to avoid harms before they can occur. Used in this way, AIAs would give government contractors that prioritize FAT in their algorithmic systems a competitive advantage in the public procurement process, incentivizing AI developers to adhere to FAT principles and practices. Likewise, companies would benefit from analyzing the impact of a proposed algorithmic system. Given the harms that internet companies' algorithmic systems can cause, stakeholders have called on those companies to conduct risk assessments before an algorithmic system is deployed.[98] An AIA framework may allow companies to evaluate their systems' impact pre-deployment and assuage their stakeholders' concerns.

Because there is little consensus on a standard AIA framework, it remains unclear how useful AIAs could be as an accountability mechanism.[99] Ultimately, their usefulness may be limited because they are merely self-assessment tools. This means potential harms discovered by a company or government agency may go unaddressed. This is especially concerning if an entity is considering deploying a high-risk algorithmic system. As a result, an entity would determine on its own what constitutes an "automated decision system" and only disclose those systems that fall under its own definition. An overly broad definition could burden companies and agencies by having to disclose irrelevant algorithmic systems, but an overly narrow definition could exclude systems that make critical and high-risk decisions about individuals' lives. Another challenge that arises surrounding AIAs in the private sector is balancing internet platforms' alleged concerns around protecting trade secrets with the goal of disclosing meaningful information on the potential impacts of algorithmic systems.[100]

Further, when public or private actors create their impact assessment framework, tensions may arise between the different values that various stakeholders want to

prioritize in evaluation practices.[101] Another limitation of AIAs is the difficulty for developers to address harms of representation in algorithmic systems (the way a system may unintentionally reinforce the marginalization of some social and cultural groups). Further, the AIA framework introduces the potential for the government and private companies' reliance on external review to become an unfunded tax on researchers and the affected communities with which they engage, who may be monitoring algorithmic decision systems without resources or compensation.[102]

Bias Impact Statements

The issue of bias in ML models, datasets, and algorithmic systems more broadly is a well-documented problem, with examples ranging from racially-discriminatory, pretrial-risk assessment tools[103] to hiring algorithms that exhibited bias against female applicants.[104] Bias can be introduced to the ML process via a number of entry points, including the use of an unrepresentative or incomplete training dataset, the use of a training dataset that reflects historical biases, poor framing of the task or fluid definitions for the model to automate, or weighting model attributes in a manner that, depending on the weights given to certain attributes, may result in bias. Researchers and advocates in the ML space have published a number of proposals to address this thorny issue, including the use of bias impact statements.

Bias impact statements are self-assessments that algorithm designers in both the government and private sector can use to evaluate the levels of bias in their model throughout the ML process. These assessments allow designers to investigate how, when, and why bias may be introduced. Bias impact statements help designers understand how the system might be biased toward certain groups and potentially inflict serious and disproportionate harm—therefore, the higher risk the algorithm, the more crucial the impact statement. Whereas AIAs offer an overall picture of an algorithmic system's impacts and harms, bias impact statements provide a focused assessment of the potential bias and discriminatory outcomes of an algorithmic model or system. Bias impact statements consist of a template of flexible questions which guide the algorithm designers' considerations while they make critical design choices throughout the development of the model. Ideally, these evaluations serve to prevent—or, at the very least, mitigate—bias that may be introduced to the model during the algorithmic design and training processes, before deployment. Also, in the event that bias is discovered during testing processes and addressed, bias impact statements also function as a historical documentation of the model's development that may be helpful in later testing and assessments. In this way, documentation throughout the ML-training life cycle is important because it allows developers to track, revisit, and understand past design decisions. It also

enables external reviewers to conduct a substantive audit of the algorithmic system.

The Brookings Institution proposed a bias impact statement framework which suggests automated decisions should be subject to scrutiny, user incentives, and stakeholder engagement.[105] The authors advise that operators of algorithms should begin by determining the possibility for unintended or negative outcomes that may result from the model and constantly question the legal, social, and economic effects and potential liabilities associated with designing the automated system. They also recommend that private companies who successfully employ bias impact statements and produce fair algorithmic outcomes be publicly acknowledged for their best practices to set an example for the rest of the industry, and that algorithm developers engage with multiple stakeholders during the design process, including civil society organizations. Civil society might also aid government and internet platforms by providing credible bias impact statement frameworks for them to use and by identifying best practices in this regard.

Bias impact statements can be helpful tools for identifying bias in algorithmic models and achieving fairer outcomes. They could provide valuable documentation of the considerations made by algorithm designers during the ML life cycle, thus aiding in holding decision makers accountable for their design choices. If made public, bias impact statements would offer a mechanism for greater transparency of AI development practices. However, bias impact statements would ultimately be a non-exhaustive self-assessment tool. There are many other factors beyond bias that designers must factor into the consideration of a model's impact, such as user privacy and safety implications. As a result, bias impact statements would have to be used in conjunction with other assessment methods. The implementation of bias impact statements is also not currently enforced, so their use is purely voluntary. Despite these limitations, however, the discovery of bias in a model is the first step toward understanding and generating solutions. In this way, bias impact statements could be one useful component of a broader solution toward promoting FAT around algorithmic systems developed by the government and internet companies.

Labels for Algorithmic Systems

Government entities and internet companies' use of algorithmic labels may help promote FAT around algorithmic systems by evaluating how effectively a system operationalizes principles such as fairness, user privacy, and safety. This quality measurement would be reflected in an algorithmic label or rating provided to consumers. A number of recent proposals have detailed frameworks for algorithmic labels or ratings as an approach to ensure FAT around algorithmic systems, and while there is variation among the specifics of each proposal (e.g.,

the methodology for assigning the rating), the proposed frameworks generally follow the same high-level model above.[106]

Algorithmic labels express how well an algorithmic system performs on a variety of indicators through an algorithmic rating that is consumer friendly. This transparency approach empowers consumers to make informed decisions regarding the technologies they use. Algorithmic labeling has been inspired by rating systems in other industries, like the EnergyStar Rating, which has become the industry standard for energy efficiency of electronic appliances, the Better Business Bureau's rating system for businesses, and the Food and Drug Administration's (FDA) nutrition label.

The notion of algorithmic labels has gained traction in the EU, where the German Data Ethics Commission has recommended a mandatory labeling scheme that would apply to public and commercial algorithmic systems—including those used by internet companies—that pose any potential risk to people's rights.[107] The obligatory labeling scheme would require operators to clearly express whether algorithmic systems are in use, and to what extent. An interdisciplinary team of experts from academic institutions in the EU, known as Bertelsmann Stiftung's AI Ethics Impact Group, has also developed a comprehensive prototype for an algorithmic labeling framework.[108] The group's 2020 working paper outlines a multi-method framework that includes an AI ethics labeling system with a rating for six key values: transparency, accountability, privacy, justice, reliability, and environmental sustainability.[109] In this framework, organizations, like government bodies and internet companies, that develop and deploy AI systems would conduct the standardized labeling process. The label is intended to account for the relevant ethical principles and to be a standardized rating that has value for all stakeholders—regulatory bodies, developers, and consumers alike.

Algorithmic labels are appealing because they feature built-in expressions of how well a model performs on important indicators in a compact standardized rating. This allows for the comparison of algorithmic systems tools via common metrics. The inherent transparency of an algorithmic label puts power into the hands of the consumers of products or platforms that employ algorithmic systems. This information enables consumers to make informed decisions about the algorithmic systems that affect their lives. These consumers have not previously been provided such critical information in a meaningful and understandable way. In the case of high-risk algorithmic systems, algorithmic labels could play a critical role in expressing the potential harms of the system to consumers, and allow consumers to avoid or otherwise mitigate harms that they could have experienced without greater understanding of how the system works.

Algorithmic labels have numerous potential benefits, but there are many difficulties involved in implementing such a system. These challenges may limit the effectiveness of labels in certain situations. Because the process of evaluating

and rating algorithms is not something that can be generalized, assessments must be focused on context-specific applications, which may result in inconsistent results. Further, the important indicators of a well-performing model will be different for each algorithm—for instance, explainability will be a much more important indicator in a public-facing algorithm than one used internally by developers. Therefore, the rating system will likely be most helpful to those who know what kind of outcome they are looking for in the algorithm in question, limiting its usefulness for everyday consumers.

We as a society must reach a consensus around which values should be prioritized in different contexts in order to construct and implement a successful algorithmic label framework. In an algorithmic labeling process, two competing values, such as privacy and transparency, may need to be balanced against each other, but determining which to prioritize is difficult and subjective. Additionally, it is currently unclear who is best positioned to function as the algorithmic labeling body, be it a current or new government entity, or an independent committee or other body established for this sole purpose. This remains a significant open question that would affect the ultimate impactfulness of algorithmic labels.

FAT Approaches Other Stakeholder Groups Can Implement

This section outlines approaches and benefits to promoting FAT around high-risk algorithmic systems that other stakeholders—such as investors, investor alliances, funders, and philanthropic organizations—can pursue.

Investor Interventions and Other External Pressures

Civil society groups, researchers, and advocates have found some success in pushing internet platforms to demonstrate transparency and accountability around their operations. But, they have thus far found less success when it comes to obtaining FAT around these companies' algorithmic systems. Investors can play a critical role in this area.

Following the attack on two mosques in Christchurch, New Zealand, the country's $41 billion sovereign wealth fund launched the first global investor coalition focused on engaging with social media companies to combat extremism and other prominent tech policy issues.[110] The coalition brought together over 100 investors with more than $13 trillion in assets to press companies to address issues related to harmful content.[111] This movement marked a scaling up in the role investors play when it comes to environmental, social, and governance (ESG) issues at technology companies. As social responsibility continues to play a growing role in business activity, investors are becoming increasingly concerned around how big tech handles issues such as data privacy, corporate surveillance, algorithmic harms, and content moderation. To this end, in 2019, the Sustainability Accounting Standards Board (SASB) launched a research project[112] to explore whether establishing standardized metrics on content moderation of internet platforms was warranted.[113] If advocates invest time in lobbying investors and standards setting organizations to pay greater attention to algorithmic harms, it could create a landscape ripe for investor pressure for greater FAT from internet platforms that use high-risk algorithmic systems. Standards-setting bodies could develop categorizations of algorithmic complexity, predictability, explainability, performance design, liability standards, best practices, and so on.[114] These standards should be developed in conjunction with civil society groups, academics, internet platforms, and other stakeholders, and should address how internet platforms develop and deploy algorithmic systems.

Investors seeking to push internet platforms for more FAT around their algorithmic systems, need to collaborate to develop clear standards related to algorithmic systems that they want companies to commit to meeting. In addition, the investor community must produce well-established accountability

mechanisms that enable them to gauge whether potential and current portfolio investments are meeting their standards around algorithmic harms. Both the standards and accountability mechanisms should be developed based on feedback and consultation with a broad range of stakeholders, including civil society groups and researchers. If widely adopted, this practice could compel emerging internet platforms to adopt responsible FAT policies around algorithmic systems, or risk losing access to capital they need to survive. If this practice is not too burdensome and becomes common among smaller companies, it could then put pressure on bigger companies that do not rely as heavily on investment sources to adopt responsible practices as well.[115]

The impact investing community has begun developing standards and accountability mechanisms that allow investors to ensure their investments are generating positive social and environmental outcomes. These mechanisms also enable investors to track whether businesses they invested in are living up to their ESG commitments.[116] Investors and investor alliances have also successfully pushed for change in corporate policies and practices related to workplace discrimination as well as climate change and resource management issues such as deforestation and water management.[117]

In addition, investors, funders, and philanthropic organizations can help establish critical structures and functions related to FAT promotion, especially when it comes to the use of high-risk algorithmic systems. For example, investors, funders, and philanthropic organizations could direct funds toward stakeholders working to increase technical expertise in auditing entities and the government, and to establish robust independent auditing mechanisms for high-risk systems[118]

These entities could also devote greater funding toward stakeholders working to train journalists, social science researchers, and civil society organizations to become more algorithmically literate, and develop educational materials to educate policymakers and the public about the harms caused by high-risk algorithmic systems and potential mitigation strategies.[119] These kinds of targeted investment or funding arrangements can help to establish a more comprehensive landscape of actors working on promoting FAT around high-risk algorithmic systems and help push the FAT agenda forward.

Recommendations

The following recommendations offer high-level guidance on how internet platforms, government entities, and policymakers can promote greater FAT around the development and deployment of high-risk algorithmic systems. In general, all efforts to promote FAT around algorithmic systems should prioritize information that is meaningful, explainable, and comprehensible by the target audience.

Recommendations for Internet Platforms

1. Provide comprehensible algorithmic system use policies that explain to consumers how the company uses algorithmic systems and for what purposes.

2. Enable users to determine whether and how their personal data is used to train a company's algorithmic systems and what data points are used to inform a company's algorithmic systems. Where possible, users should also be able to opt out of having algorithmic systems shape their online experiences.

3. Expand transparency efforts to include more quantitative and qualitative information on how platforms use algorithmic tools to deliver their services. Where relevant, this should include more information on how companies use algorithmic systems to moderate content and to target and deliver ads, what the error rates of these systems are, and what impact these systems have had on user speech and experiences.

4. Submit to regular independent external audits and commit to, at a minimum, publishing a public summary of the findings and making subsequent mitigation efforts.

5. Supplement potential external auditing efforts by conducting proactive, regular internal audits of algorithmic systems in order to identify potentially harmful outcomes related to privacy, freedom of expression, freedom of information, or cases of discrimination surfaced by community partners, civil society organizations, activists, etc. Companies should take steps to eliminate or mitigate any harms identified. Companies should also share summaries of the audits in a public and explainable manner.

6. Collaborate with civil society, researchers, and government agencies to create standardized mechanisms for benchmarking ML documentation procedures. At a minimum, these efforts should include the creation of documentation procedures for datasets on which models are trained, intended use cases of models, and the performance characteristics of models. Ideally, documentation procedures will be flexible enough to accommodate the model-specific needs of an AI system while still following a standardized format, and strategically communicate information so that it is valuable to technical and non-technical stakeholders alike.

7. Establish mechanisms for consumers to provide feedback on adverse outcomes that result from an algorithmic system, such as the appeals process currently available to users subject to content moderation processes. This will encourage greater company accountability to their consumers.

Recommendations for Government Entities and Policymakers

1. Pass comprehensive federal privacy legislation that requires internet platforms to provide transparency, impact assessments, and regular audits to prevent algorithmic tools from being used in ways that disproportionately impact disadvantaged communities. These rules should also empower the FTC or a Data Protection Authority to enforce requirements.

2. Require government agencies that develop and/or deploy algorithmic systems to conduct periodic algorithmic audits and impact assessments to identify and mitigate discrimination, bias, and other harms. One way to achieve this would be for the current administration to introduce an EO requiring government agencies to evaluate any high-risk algorithmic systems pre-deployment. These systems should also be subject to continuous and periodic reviews to account for changes in systems and how they operate. The EO should establish a clear definition for high-risk systems to ensure that any legal actions are proportionate. This definition should be developed based on multi-stakeholder consultations and dialogue, broadly applicable, and dynamic enough to account for the fact that the risks an algorithmic system poses can change over time. Government agencies should also conduct regular impact assessments and algorithmic audits for high-risk algorithmic systems, even when not required, to identify and mitigate discrimination, bias, and other potential harms.

3. Supplement the EO discussed above with clear rules that require companies and government agencies to review their algorithmic systems, particularly their high-risk algorithmic systems, before they are deployed and to mitigate any identified harms. If these entities fail to do so, they can be held liable by a regulator.

4. Task government agencies, particularly NIST, with establishing a robust set of auditing standards for internet platform use of algorithmic systems. In practice, these standards should result in a distinct professional auditing field that conducts external, independent audits on technology companies' use of algorithmic systems. Similar to the EO, auditing standards should establish a clear definition of "high-risk systems" based on multi-stakeholder consultations and dialogue. This definition should be broadly applicable and should be dynamic enough to account for the fact that the risks an algorithmic system poses can change over time.

5. Establish a set of safeguards that enable internet platforms that develop and deploy high-risk algorithmic systems to share data with vetted researchers and auditors in a manner that mitigates privacy and trade secret concerns and facilitates meaningful research and evaluation. Internet platforms should also avoid using the CFAA as a mechanism for penalizing researchers and auditors who are working to promote FAT around algorithmic systems.

6. Institute mechanisms, incentives, and programs for recruiting and upskilling technical talent to staff government agencies so AI tools can be developed internally, or at least overseen and evaluated for issues by in-house staff.

As conversations around promoting FAT around high-risk algorithmic systems continue, we also recommend that a broad range of stakeholders—including internet platforms, government entities, civil society organizations, researchers, investors, and funders—come together to pursue the following recommendations. These recommendations encourage collaborative efforts with the aim of developing meaningful and actionable standards related to high-risk algorithmic systems.

1. Develop a clear set of standards and methodologies to guide the implementation of algorithmic audits for internet platforms. These guidelines should include details on procedure, transparency, and strategies to mitigate harms. While these guidelines should offer a degree of standardization to auditors, they should also account for variances between algorithmic systems and harms they can cause and leave room for these unique characteristics to be considered.

2. Collaborate with standards setting bodies to establish clear standards around how companies can develop and deploy algorithmic systems and clear ESG commitments related to use of these systems. These standards should be accompanied by mechanisms that enable investors to evaluate whether companies are fulfilling their commitments. Standards-setting bodies should also establish mechanisms for soliciting feedback on standards for a broad range of stakeholders, including civil society groups and researchers.

3. Direct funding and resources toward efforts seeking to establish a robust landscape of actors working on FAT around algorithmic systems. This includes funding for robust technical training for civil society organizations, journalists, government agencies, and auditing entities, as well as awareness and education events for policymakers and the public.

Conclusion

Internet platforms and government agencies are rapidly expanding their development and use of ML and AI-based technologies. As this report outlines, there are a broad range of mechanisms that these entities can use to promote FAT around high-risk algorithmic systems. Currently, however, efforts to promote FAT around high-risk algorithmic systems tend to focus on a selection of mechanisms and do not consider how these mechanisms can be deployed in concert with one another. Going forward, internet platforms, government entities, and other relevant stakeholders must work to develop comprehensive policies, standards, and roadmaps that can generate meaningful FAT around high-risk algorithmic systems.

Notes

1 Tom Simonite, "Meet the Secret Algorithm That's Keeping Students Out of College," *WIRED*, July 2020, https://www.wired.com/story/algorithm-set-students-grades-altered-futures/.

2 Drew Harwell, "A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job," *Washington Post*, November 6, 2019, https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/.

3 Alex Chohlas-Wood, *Understanding Risk Assessment Instruments in Criminal Justice*, June 19, 2020, https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice/.

4 Muhammad Ali et al., "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes," *Proceedings of the ACM on Human-Computer Interaction 2019*, September 12, 2019, https://arxiv.org/abs/1904.02095.

5 Christine Bannan and Margerite Blase, *Automated Intrusion, Systemic Discrimination: How Untethered Algorithms Harm Privacy and Civil Rights*, October 7, 2020, https://www.newamerica.org/oti/reports/automated-intrusion-systemic-discrimination/.

6 Brent Mittelstadt. 2019. *AI Ethics: Too Principled to Fail?* SSRN (2019).

7 David Freeman Engstrom et al., *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*, February 2020, https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

8 Spandana Singh, *Holding Platforms Accountable: Online Speech in the Age of Algorithms*, https://www.newamerica.org/oti/reports/report-series-content-shaping-modern-era/.

9 Margaret Mitchell et al., "Model Cards for Model Reporting," *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29--31, 2019, Atlanta, GA, USA*, January 14, 2019, https://arxiv.org/abs/1810.03993.

10 Mitchell et al., "Model Cards".

11 “Datasheets for Datasets,” Microsoft Research, March, 23, 2018, https://www.microsoft.com/en-us/research/publication/datasheets-for-datasets/.

12 “AI FactSheets 360,” IBM Research, https://aifs360.mybluemix.net/.

13 Mitchell et al., "Model Cards".

14 Mitchell et al., "Model Cards".

15 Mitchell et al., "Model Cards".

16 Isabel Kloumann and Jonathan Tannen, "How We're Using Fairness Flow to Help Build AI That Works Better for Everyone," Facebook AI, last modified March 31, 2021, https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/.

17 Mitchell et al., "Model Cards".

18 Spandana Singh, one of the authors of this report, represents New America on the ABOUT ML Steering Committee.

19 Kevin Bankston and Ross Schulman, *Getting Internet Companies To Do The Right Thing*, February 9, 2017, https://www.newamerica.org/in-depth/getting-internet-companies-do-right-thing/case-study-3-transparency-reporting/.

20 Spandana Singh and Kevin Bankston, *The Transparency Reporting Toolkit: Content Takedown Reporting*, October 25, 2018, https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/.

21 Spandana Singh and Leila Doty, *The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules*, April 8, 2021, https://www.newamerica.org/oti/reports/transparency-report-tracking-tool/.

22 Facebook, *Community Standards Enforcement Report*, https://transparency.fb.com/data/community-standards-enforcement/?from=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement.

23 Google, *YouTube Community Guidelines Enforcement*, https://transparencyreport.google.com/youtube-policy/removals?hl=en.

24 Ranking Digital Rights, "The 2020 RDR Index," Ranking Digital Rights, https://rankingdigitalrights.org/index2020/.

25 Singh and Doty, *The Transparency*.

26 Mike Isaac and Daisuke Wakabayashi, "Russian Influence Reached 126 Million Through Facebook Alone," *New York Times*, October 30, 2017, https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html.

27 Aaron Rieke and Corrine Yu, "Discrimination's Digital Frontier," *The Atlantic*, April 15, 2019, https://www.theatlantic.com/ideas/archive/2019/04/facebook-targeted-marketing-perpetuates-discrimination/587059/. Spandana Singh, *Special Delivery: How Internet Platforms Use Artificial Intelligence to Target and Deliver Ads*, February 18, 2020, https://www.newamerica.org/oti/reports/special-delivery/.

28 Facebook, "Ad Library," Facebook, https://www.facebook.com/ads/library.

29 Google, *Political Advertising on Google*, https://transparencyreport.google.com/political-ads/home.

30 Reddit, "Reddit Political Ads Transparency Community," Reddit, https://www.reddit.com/r/RedditPoliticalAds/.

31 Singh, *Special Delivery*. Spandana Singh, "Reddit's Intriguing Approach to Political Advertising Transparency," *Slate's FutureTense*, May 1, 2020, https://slate.com/technology/2020/05/reddit-political-advertising-transparency.html.

32 https://www.facebook.com/ads/manage/customaudiences/tos.php. This is a customized settings page that is available to each logged in user.

33 Aaron Rieke and Miranda Bogen, *Leveling the Platform: Real Transparency for Paid Messages on Facebook*, May 2018, https://www.upturn.org/reports/2018/facebook-ads/.

34 Santa Clara Principles on Transparency and Accountability in Content Moderation, https://santaclaraprinciples.org/.

35 Ranking Digital Rights, "The 2020," Ranking Digital Rights.

36 Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, July 22, 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/.

37 European Commission, *Proposal for a Regulation of the European Parliament And of the Council on a Single Market For Digital Services (Digital Services Act) and Amending Directive 2000/31/EC*, December 15,
2020, https://ec.europa.eu/info/sites/default/files/proposal_for_a_regulation_on_a_single_market_for_digital_services.pdf.

38 Platform Accountability and Consumer Transparency Act, S. 4066, 116th, 1st Sess. https://

www.congress.gov/bill/116th-congress/senate-bill/4066/text.

39 European Parliamentary Research Service, *A Governance Framework for Algorithmic Accountability and Transparency*, April 2019, https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf.

40 Santa Clara Principles on Transparency and Accountability in Content Moderation.Singh, *Everything in Moderation*.

41 European Commission, *White Paper on Artificial Intelligence - A European Approach to Excellence and Trust*, February 19, 2020, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

42 European Commission, *White Paper*.

43 European Commission, *White Paper*.

44 European Parliamentary Research Service, *A Governance*.

45 European Parliamentary Research Service, *A Governance*.

46 Engstrom et al., *Government by Algorithm*.

47 European Parliamentary Research Service, *A Governance*.

48 Engstrom et al., *Government by Algorithm*.

49 Andrew Tutt, "An FDA for Algorithms," *Administrative Law Review* 69, no. 83 (March 2016): http://dx.doi.org/10.2139/ssrn.2747994.

50 European Parliamentary Research Service, *A Governance*.

51 Matthew U. Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law & Technology* 29, no. 2 (Spring 2016): https://dx.doi.org/10.2139/ssrn.2609777.

52 European Commission, *White Paper.*

53 In the EU, this approach would also require coordination with relevant data protection authorities.

54 European Parliamentary Research Service, *A Governance*.

55 European Parliamentary Research Service, *A Governance*.

56 "Solar Investment Tax Credit (ITC)," Solar Energy Industries Association, https://www.seia.org/initiatives/solar-investment-tax-credit-itc. Congressional Research Service, *The Renewable Electricity Production Tax Credit: In Brief*, April 29, 2020, https://fas.org/sgp/crs/misc/R43453.pdf.

57 European Parliamentary Research Service, *A Governance*.

58 “Germany: Flawed Social Media Law,” Human Rights Watch, February 14, 2018, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

59 European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, April 4, 2021, https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF.

60 European Commission, *Proposal for a Regulation*.

61 Spandana Singh, "The EU's Digital Services Act Makes a Positive Step Towards Transparency and Accountability, But Also Raises Some Serious Questions," New America's Open Technology

Institute, last modified January 21, 2021, https://www.newamerica.org/oti/blog/the-eus-digital-services-act-makes-a-positive-step-towards-transparency-and-accountability-but-also-raises-some-serious-questions/.

62 Spandana Singh, "Breaking Down the World's First Proposal for Regulating Artificial Intelligence," New America's Open Technology Institute, last modified June 10, 2021, https://www.newamerica.org/oti/blog/breaking-down-the-worlds-first-proposal-for-regulating-artificial-intelligence/.

63 Exec. Order No. 13859 Fed. Reg. (Feb. 14, 2019). https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.

64 Exec. Order No. 13859

65 "AI Standards: Federal Engagement," National Institute of Standards and Technology, March 14, 2019, https://www.google.com/url?q=https://www.nist.gov/artificial-intelligence/ai-standards&sa=D&source=editors&ust=1628792098629484&usg=AOvVaw3qNw7gwlNZRdJOGVgSUgYK.

66 Exec. Order No. 13960 Fed. Reg. (Dec. 8, 2020). https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.

67 Exec. Order No. 13960

68 Bannan and Blase, *Automated Intrusion*.

69 Bannan and Blase, *Automated Intrusion*.

70 Leila Doty and Lauren Sarkesian, "To Ensure More Trustworthy AI, Use an Old Government Tool: Public Procurement," *Issues in Science and Technology*, February 9, 2021. https://issues.org/trustworthy-ai-federal-procurement-reform/.

71 "A Snapshot of Government-wide Contracting for FY 2019 (infographic)," U.S. Government Accountability Office Watchblog, May 26, 2020, https://blog.gao.gov/2020/05/26/a-snapshot-of-government-wide-contracting-for-fy-2019-infographic/

72 "Government Procurement: What is Government Procurement?," FindRFP, https://www.findrfp.com/Government-Contracting/Government-procurement.aspx#:~:text=Government%20procurement%2C%20also%20known%20as,goods%20and%20services%20they%20provide. Congressional Research Service, *Defense Primer: Lowest Price Technically Acceptable Contracts*, January 22, 2021, https://fas.org/sgp/crs/natsec/IF10968.pdf.

73 Doty and Sarkesian, "To Ensure".

74 "Policy and Investment Recommendations for Trustworthy Artificial Intelligence," European Commission, June 26, 2019 https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

75 "Grip on Algorithms," Township Amsterdam, https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/grip-op-algoritmes/

76 Engstrom et al., *Government by Algorithm*.

77 For instance, private companies may withhold information that is not necessary to the bidding suppliers, they may choose which vendors they request proposals from, and they are not required to publish their contract awards in the same way that public entities are required."Private Vs. Public Sector Procurement Practices," Concord, last modified April 3, 2019, https://www.concordnow.com/blog/private-vs-public-sector-procurement-practices/. "Private vs. Public Sector Bidding Process," Handex, last modified January 8, 2019, https://www.hcr-llc.com/blog/private-vs.-public-bidding-process.

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

79 Inioluwa Deborah Raji et al., "FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency," *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, January 2020, https://arxiv.org/pdf/2001.00973.pdf.

80 Howard B. Levy, "History of the Auditing World, Part 1," *The CPA Journal*, https://www.cpajournal.com/2020/11/25/history-of-the-auditing-world-part-1/.

81 Raji et al., "FAT* '20".

82 Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, May 22, 2014, http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

83 Engler, "Independent Auditors".

84 Raji et al., "FAT* '20".

85 Raji et al., "FAT* '20".

86 The CFAA prohibits accessing a computer without authorization in order to curb hacking, but does not define "authorized access," leaving it to website operators to determine, and thwarting research and other legitimate access in the process.

87 James Guszcza et al., "Why We Need to Audit Algorithms," *Harvard Business Review*, November 28, 2018, https://hbr.org/2018/11/why-we-need-to-audit-algorithms.

88 Guszcza et al., "Why We Need".

89 Guszcza et al., "Why We Need".

90 Raji et al., "FAT* '20".

91 Guszcza et al., "Why We Need".

92 AI Now Institute, *Confronting Black Boxes: A Shadow Report of the New York City Automated Decision System Task Force*, December 4, 2019, https://ainowinstitute.org/ads-shadowreport-2019.html.

93 AI Now Institute, *"Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability",* April 2018, https://ainowinstitute.org/aiareport2018.pdf.

94 "H.R.2231 - Algorithmic Accountability Act of 2019," The United States Congress, April 2019, https://www.congress.gov/bill/116th-congress/house-bill/2231.

95 European Parliamentary Research Service, *A Governance*.

96 Critically, AI Now's AIA structure allots a due process challenge period, which provides a path for the public to challenge the adoption of the automated system if the agency fails to comply with AIA requirements or performs a substandard self-assessment.

97 Bannan and Blase, *Automated Intrusion*.

98 Ranking Digital Rights, "2019 RDR Index Methodology," Ranking Digital Rights, 2019, https://rankingdigitalrights.org/index2019/report/index-methodology/.

99 "Governing with Algorithmic Impact Assessments: Six Observations," Data & Society, April 24, 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3584818.

100 Rieke and Bogen, *Leveling the Platform*.

101 One well-known example of an impact assessment framework is a human rights impact assessment (HRIA), which draws from human rights principles, such as those on freedom of expression and privacy, and can be used to measure the impact that technologies have on individuals' fundamental rights. HRIAs are a valuable mechanism for assessing FAT in a human rights context. However, a persistent challenge related to conducting an HRIA-like evaluation is that these assessments require a large amount of data, some of which may not be disclosed publicly by companies and some of which may be sensitive personal information. It is unclear how to best address this fundamental tension between wanting to implement impact assessments on groups that may be the most harmed by technologies and wanting to uphold strong privacy standards by not collecting huge datasets, which may include sensitive information that could be weaponized against said groups. Ranking Digital Rights, "2019 RDR Index," Ranking Digital Rights.Nora Götzmann, ed., *Handbook on Human Rights Impact Assessment* (Edward Elgar Publishing, 2019).

102 AI Now Institute, *Confronting Black.*

103 "Machine Bias," ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

104 "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," Reuters, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

105 "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms," Brookings Institution, May 22, 2019, https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

106 "Could Rating Systems Encourage Responsible AI?," Brookings Institution, October 1, 2020, https://www.brookings.edu/events/could-rating-systems-encourage-responsible-ai/.

107 In 2018, the Chancellor of Germany, Angela Merkel, tasked the German Data Ethics Commission with producing recommendations for rules around AI to protect individual rights, preserve social cohesion, and safeguard and promote prosperity in the information age. The report recommended that algorithmic systems should be designed to protect democracy and people's rights and freedoms, be secure, and avoid bias and discrimination."Data Ethics Commission," Federal Ministry of the Interior, Building, and Community, September 2018, https://www.bmi.bund.de/EN/topics/it-internet-policy/data-ethics-commission/data-ethics-commission-node.html."Opinion of the Data Ethics Commission," Daten Ethik Kommission, October 2019, https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2.

108 "From Principles to Practice: How Can We Make AI Ethics Measurable?," Ethics of Algorithms, April 2, 2020, https://ethicsofalgorithms.org/2020/04/02/from-principles-to-practice-how-can-we-make-ai-ethics-measurable/.

109 The AI ethics label is supplemented by a risk matrix and VCIO (Values, Criteria, Indicators, Observables) (VCIO) model. To reach a specific rating level for a given value, the minimum requirements of observable factors must be met (e.g., to receive an 'A' rating for privacy, end-to-end encryption must be a feature of the AI system). The value ratings of the ethics label are underpinned by the VCIO model, which identifies a key AI ethics value and specifies criteria that define the fulfillment or violation of that value, acknowledging that values are often in conflict with each other and proposing a process for hierarchizing values against one other when this is the case. The framework's risk matrix accounts for the context-dependent nature of AI applications with its two-dimensional classification system.

110 Gillian Tett, "ESG Investors Are Taking on Big Tech," *Financial Times*, December 19, 2019, https://www.ft.com/content/adee5360-2265-11ea-b8a1-584213ee7b2b.

111 Tett, "ESG Investors".

112 "Content Moderation on Internet Platforms," Sustainability Accounting Standards Board, https://www.sasb.org/standards/process/active-projects/content-moderation-on-internet-platforms-research-project/.

113 SASB is an independent nonprofit organization that creates standards for company disclosures on ESG issues that are most relevant to financial performance to their investors.

114 European Parliamentary Research Service, *A Governance*.

115 Sean Collins and Kristen Sullivan, "Advancing Environmental, Social, and Governance Investing," Deloitte Insights, https://www2.deloitte.com/us/en/insights/industry/financial-services/esg-investing-performance.html.

116 Sophie Edwards, "Impact Investors Must Set Up 'Accountability Tools,' Experts Say," Devex, last modified April 13, 2018, https://www.devex.com/news/impact-investors-must-set-up-accountability-tools-experts-say-92528. "What You Need to Know About Impact Investing," Global Impact Investing Network, https://thegiin.org/impact-investing/need-to-know/.

117 Ceres, Environmental Defense Fund, and KKS Advisors, *The Role of Investors in Supporting Better Corporate ESG Performance*, February 2019, https://www.ceres.org/sites/default/files/reports/2019-04/Investor_Influence_report.pdf. A whole body of literature exists in the ESG space around how to effectively lobby investors to press for change and how investors can make the biggest impact.

118 European Parliamentary Research Service, *A Governance*.

119 AI Now Institute, *Confronting Black*.

NEW AMERICA