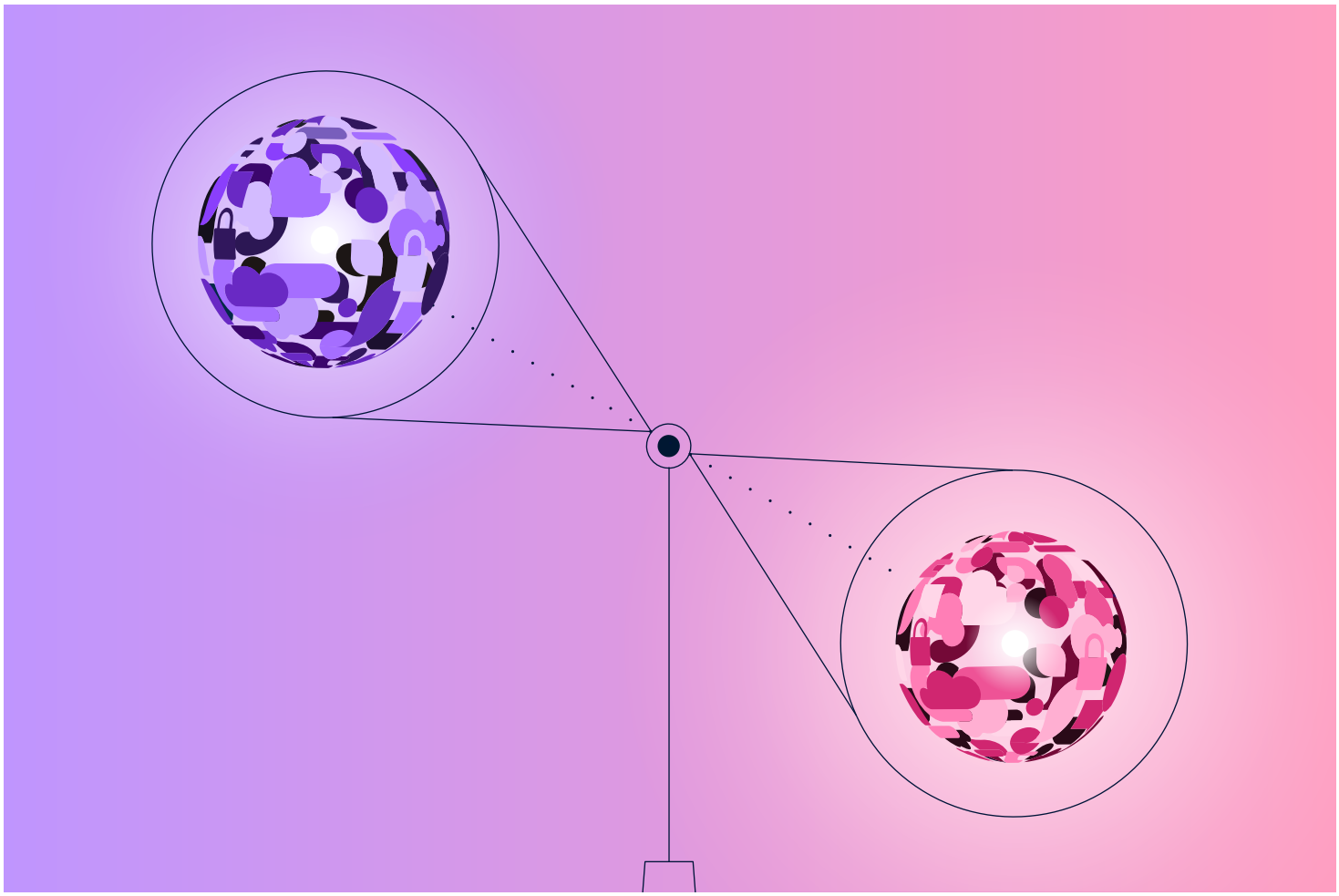
Standards for protecting at-risk groups in AI bias auditing

November 2022



IBM

Table of contents

The use of artificial intelligence (AI) systems is growing rapidly, with a worldwide market that is expected to reach nearly $450 billion in 2022 and nearly $900 billion in 2026[1]. Technical advancements and competitive forces are driving adoption, making the use of AI a potential differentiator for businesses in the current digital economy.

AI systems are used to make predictions, recommendations, or decisions and their outputs or behaviors are not necessarily pre-determined by their developer or user. This makes these systems highly useful given their ability to augment human efforts and drive efficiency. However, like the humans they help, these systems can also have the potential to contain unfair bias, particularly for the members of at-risk groups of individuals who might not be well represented in the training data sets or fully considered in the design of these systems.

The risk of bias

AI has many benefits when used appropriately and can even help mitigate against human bias in some cases. However, bias in AI is a widely acknowledged risk, with high-profile cases regularly making the news headlines. Bias can be introduced into AI systems, intentionally or not, because of cultural, technical, or deployment factors. Bias can be thought of as a systematic error that may potentially cause the AI system to generate unfair decisions.

Recently, concerns around privacy and bias led the U.S. Internal Revenue Service to back away from using facial recognition in taxpayer identity verification[2]. Other widely known cases involving issues of bias include AI systems used in recidivism rate estimation, patient health risk estimation, and hiring.

Consider the case of hiring for an open position at a company, and how bias can be inadvertently built into the process. A company might use an AI system for resume/CV screening that recommends the candidates most likely to be hired by the company. If the company has historically hired more white males than members of other demographic groups and then uses their historical data as input to train the AI system to identify the candidates most likely to be hired, then the output from that AI system may also be biased towards recommending more white males as candidates for hiring.

Policymaking and bias audits

Governments around the world are discussing and developing policies to address potential bias in AI systems. This includes the development of requirements like bias auditing, particularly in high-risk systems.

New York City has recently enacted Local Law 144 regarding automated employment decision tools that will go into effect on January 1, 2023. This law requires an independent bias audit be conducted on all AI systems used to screen residents for employment or promotion decisions.

Other states and localities have also introduced proposals that include bias audit requirements. For example, in Washington, DC, the proposed bias audit requirement would apply not just to employment related decisions – like those where U.S. Equal Employment Opportunity Commission (EEOC) guidance applies – but also to other types of AI systems including those used in housing and public services where there is no similar EEOC guidance[3]. In October 2021, the EEOC launched an initiative focused on AI and algorithmic fairness that aims to support compliance with anti-discrimination laws[4]. The EEOC also recently released a new technical assistance document regarding the Americans with Disabilities Act and the Use of Software, Algorithms, and AI to Access Job Applicants and Employees[5].

In addition, other government organizations within the U.S., such as the Federal Trade Commission, have also signaled their intent to enforce protections against potential bias in AI, and it is reasonable to expect that such enforcement might also include requirements for bias auditing[6]. Signals that bias audits might be a focus for AI going forward are found in published content from governments in other regions around the globe as well.

For example, the government of Singapore has published a governance framework for AI that includes consideration for algorithmic audits in certain circumstances[7]. In Europe, the Charter of Fundamental Rights of the EU and the European Convention on Human Rights provide the human rights framework for the use of AI[8]. The Charter of Fundamental Rights of the EU prohibits unfair bias on the grounds of protected characteristics[8].

The European Commission (EC) also has both existing anti-discrimination directives as well as proposed AI regulation that could contribute to the increased use of bias audits[9]. The global push for AI regulation and bias audits is likely to make bias auditing common practice.

Current challenges associated with bias audits

Bias audits, while positively intentioned, do not come without their own risks and challenges. These audits can be done both as internal assessments and by third-party external auditors. Auditors of AI systems will be required to make decisions about how to translate legal standards into practice. Today, we lack consensus on how this will be done, and the relevant laws are not always consistent.

For example, the New York City law focuses on bias assessments for the protected characteristics of ethnicity, race and gender, while other proposed legislation references longer lists that includes other characteristics such as familial status and source of income.

> The lack of consistency will be challenging as auditors and AI developers decide how to develop processes and tooling for bias detection and mitigation.

Auditors and developers will also be faced with decisions about how to group the underlying values of the AI system's data points that relate to the protected characteristics. Protected characteristics are attributes associated with an individual that can be the basis for unfair social bias (e.g., race, gender, age, disability).

The data points associated with these protected characteristics can be grouped into classes; for example, gender as a protected characteristic would include, at minimum, the classes of male and female. Auditing for bias requires comparing AI system results for members of at least two protected classes of an associated characteristic.

For example, to assess for gender bias, one might assess whether males tended to receive more favorable results than females.

In an algorithm intended to screen candidates for hiring decisions, the system would be considered to have an unfair bias if it systematically screened out significantly more females than males from the potential hiring pool when the male and female populations are similar with respect to relevant hiring criteria. To do this assessment for bias, auditors and developers need to know which protected characteristics and associated classes they need to consider and then they will make decisions about which data points represent these classes.

Bias audits also present challenges related to the highly nuanced issue of bias itself. For example, within a hiring context it is important to fill open positions with the right level of skills and experience and the candidate pool is not always equally balanced across protected classes.

Data associated with protected characteristics might not even be collected in certain countries and contexts due to regulatory restrictions on data collection due to historical inequities. If the data is available, developers might choose to exclude protected characteristics from their models due to concern about introducing bias; however, proxies for these characteristics can still introduce bias. For example, browsing history content can be tied to gender, and zip code information can be tied to race[10].

Complicating the issue of whether to use protected characteristic data is the fact that the use of this data to debias AI models can introduce risk in highly regulated environments where disparate treatment is a concern, such as in a hiring context[10].

Additionally, focusing only on bias for at-risk groups has the potential to raise concerns of inadvertently creating bias for other groups. Bias audit tools tend to focus only on detecting bias for at-risk groups by comparing them to a reference group which is assumed to have a lower risk of unfair bias.

Defining protected class data

Defining protected class data might seem simple; however, there is widespread debate about treating characteristics such as gender as a binary data point. Racial categories are also debated as social constructs that cannot easily be reduced to a single data point.

It might be clear how to define classes for age in certain contexts, but less clear in others. For example, if the AI is used in employment decisions within the United States, then it could make sense to group ages into two classes of over 40 years old and under 40 years old based on the Age Discrimination in Employment Act, which is enforced by the U.S. Equal Employment Opportunity Commission[11]. However, determining the age classes to test for unfair bias in AI systems used in other contexts could be less clear and, therefore, open for more debate.

Defining protected class data is the most ethically concerning issue.

Social bias that is already present can seep into auditing practices if they do not even consider certain classes, making it impossible to protect individuals within those classes. You cannot find what you do not look for. There cannot be impactful bias auditing, assessing whether bias against a protected class is present if the auditor or developer is not running checks for the data points associated with those protected classes.

For example, if bias is only assessed for age classes of over 40 years old and under 40 years old, then any bias for the age class of less than 12 years old could go undetected if the bias is masked by being included in the broader class of under 40 years old.

Regional differences, including cultural and legal differences, and use case are also important to consider when evaluating which protected characteristics and classes to consider in assessing systems for bias. For example, while race can be used as one important reference characteristic when assessing for bias within the United States, in another country it might be appropriate to put increased focus on ethnicity as a protected characteristic and consider individuals who identify with the largest ethnic group as a key reference class in bias auditing.

Implications

The implications of these challenges depend on the context of where and how an AI system is used, but clearly without adequate guidance there are significant risks introduced by bias auditing. The first risk cuts to the core of the very intent of bias auditing, which is to detect bias. This happens because the bias can go undetected for individuals associated with protected characteristics or classes that are not included in standard bias auditing practices. The second risk is that successfully passing a bias audit will create confidence in an AI system that has been assessed for bias only for a subset of the possible protected characteristics and classes.

> It should be standard practice in bias audit reporting to articulate the assumptions used for determining the relevant protected characteristics and associated classes used in the bias audit.

Auditors and developers need guidance to know which protected characteristics to test for and how to group the protected class data. In practice, developers may be more likely to avoid bias testing when the guidance is unclear. Developers also need to evaluate whether training or input data reflects cultural bias.

For example, without clear guidance, developers might unintentionally introduce bias into the AI system if they attempt to debias the system using data that reflects cultural bias. With regulations forthcoming around external bias auditing practices, a lack of guidance potentially leaves the operators of AI systems exposed to the review of external auditors without a clear understanding of how to run internal audits and address potential bias in the systems they develop and use.

Adaptive tooling for bias audits

While auditors and developers will require guidance and standards to conduct consistent bias audits on AI systems, protected characteristics and their associated classes will remain a matter of local and legal debate. There will always be some level of discretion across contexts and locations. It is also important that developers of AI be able to reflect their own ethical standards when evaluating AI systems for bias, which may go beyond regulatory requirements. Therefore, the tooling that auditors and developers use to conduct bias testing should remain flexible and allow users to easily adapt testing to local norms, stakeholder requirements, and legal expectations for the location and context in which the system is deployed.

Given the potential differences in relevant testing criteria, the ability to indicate during the audit process that a standard protected characteristics or class is not relevant to the AI system would also be important. As the foundation of the auditing industry is built out over the next several years, building in flexibility through adaptive tooling is essential; however, there is evidence that current auditing tools may not contain this type of flexibility.

Current open-source toolkits for the measurement of bias that specifically use the term “audit” tend to either use a pre-determined list of protected class data points and groupings or assume data points and groupings from the dataset itself. Other open-source toolkits that enable the measurement of bias within AI systems may provide similar capabilities and additional flexibility, but do not typically highlight the term “audit”.

It is likely that as more regulations are passed requiring bias audits there will be an increased demand for tools specifically designed to support them. The industry needs adaptive bias auditing tools, with flexibility built-in by design.

Moving Forward

The use of AI throughout society can present many benefits, fueling its rapid adoption throughout society. These benefits can be enhanced through actions that may enable trust in fair outcomes, including the development of standards relating to bias audits.

These standards can inform auditors and developers of AI on what protected characteristics should be considered in bias audits and how to translate those into data points required to conduct these assessments. Given that the bias auditing industry is likely to grow rapidly over the next few years, it is important to think carefully about the assumptions regarding protected characteristics that may be built into bias audit tooling.

We welcome continued conversations with stakeholders and believe there is a need for the consideration of diverse perspectives to adequately address the associated ethical concerns and help enable best practices. This ongoing dialogue will be of critical importance as the world begins creating the foundational structure for how bias auditing is done moving forward.

Acknowledgement

The authors would like to thank the IBM AI Ethics Board members, Milena Pribic, and the other subject matter experts who reviewed and provided their feedback on this document.

Authors

Heather Domin
Program Director, AI Governance

Jamie VanDodick
Director, Tech Ethics Project Office and Governance

Calvin Lawrence
Distinguished Engineer, Chief Architect Cognitive Solutions & Innovation (AI) Public Sector

Francesca Rossi
IBM fellow and AI Ethics Global Leader

References

1. IDC Semiannual Artificial Intelligence Tracker, 2H 2021, July 2022
2. IRS announces transition away from use of third-party verification involving facial recognition, IRS, irs.gov, 7 February 2022
3. AG Racine Introduces Legislation to Stop Discrimination In Automated Decision-Making Tools That Impact Individuals' Daily Lives, OAG DC, oag.dc.gov, December 9, 2021
4. EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness, EEOC, eeoc.gov, 28 October 2021
5. The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees, EEOC, eeoc.gov, 12 May, 2022
6. Aiming for truth, fairness, and equity in your company's use of AI, Elisa Jillson. ftc.gov, 19 April 2021
7. Modal Artificial Intelligence Governance Framework, Info-communications Media Development Authority (IMDA) and Personal Data Protection Commission (PDPC), 2020
8. Getting the future right, artificial intelligence ad fudamental rights, European Union Agency for Fundamental Rights, 2020
9. Proposal for a regulation of the European parlliament and of the council laying down harmoised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, European Commission, April 2021
10. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST, March 2022
11. Age Discrimination, EEOC, eeoc.gov

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
November