Hearing on

"Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age"

United States House of Representatives Committee on Financial Services
Task Force on Artificial Intelligence

October 13, 2021, at 12:00 p.m.

Testimony of Aaron Cooper
Vice President, Global Policy
BSA | The Software Alliance

Testimony of Aaron Cooper
Vice President, Global Policy, BSA | The Software Alliance
Hearing on "Beyond I, Robot: Ethics, Artificial Intelligence, and the Digital Age"

Before the United States House of Representatives
Committee on Financial Services
Task Force on Artificial Intelligence

October 13, 2021

Good afternoon Chairman Foster, Ranking Member Gonzalez, and members of the AI Task Force. My name is Aaron Cooper. I am Vice President of Global Policy for BSA | The Software Alliance (BSA).

BSA is the leading advocate for the global software industry.[1] Our members are at the forefront of developing cutting-edge, data-driven services that have a significant impact on US job creation and growing the global economy. I commend the Task Force for convening today's important hearing, and I thank you for the opportunity to testify.

Enterprise software services, including artificial intelligence (AI) are accelerating digital transformation in every sector of the economy. Artificial intelligence is not just about robots, self-driving vehicles, or social media. It can be used by businesses of all sizes to improve their competitiveness, enhance their value proposition, and increase their capacity to make data-informed decisions.

BSA represents the perspective of the enterprise software companies that help make this possible. Our members create the technology products and services that help other businesses innovate and grow. In that capacity, BSA members are on the leading edge of providing businesses in every sector of the economy with the trusted tools they need to leverage the benefits of AI.

The promise that AI may one day impact every industry is quickly turning into a commercial reality and driving the digital transformation. For instance, Autodesk brings the power of AI to industrial design, helping American manufacturers improve the performance of their products while reducing their costs and environmental impact. In one recent collaboration, Autodesk worked closely with engineers at General Motors to explore how AI-enabled generative design could help the company optimize its manufacturing processes.[2] As an initial proof-of-concept, the two companies set out to improve GM's approach to designing and manufacturing the brackets that secure seatbelts and seats to a car's floor. The partnership yielded immediate benefits, enabling GM to identify a new design that is 40 percent lighter and 20 percent stronger than its previous approach.

---

[1] BSA's members include: Adobe, Atlassian, Autodesk, Bentley Systems, Box, CNC/Mastercam, DocuSign, IBM, Informatica, MathWorks, Microsoft, Okta, Oracle, PTC, Salesforce, ServiceNow, Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, Twilio, Workday, Zendesk, and Zoom Video Communications, Inc.

[2] General Motors | Generative Design in Car Manufacturing | Autodesk

Splunk is helping the financial services sector leverage AI to take a bite out of the more than $40 billion that is lost to fraudulent transactions each year. Splunk's software powers a suite of enterprise fraud management capabilities that allow banks to identify transaction anomalies in real time, reduce the frequency of false positives, and better protect consumers from identity theft.[3]

While the adoption of AI can unquestionably be a force for good, it can also create real risks if not developed and deployed responsibly. We commend the Task Force for convening today's hearing to examine the role that frameworks for ethical AI can play in ensuring the responsible use of this technology. This is an area of particular focus for BSA and our member companies are leaders when it comes to responsible AI practices.[4] We recently produced a detailed framework that sets forth a risk management approach for confronting concerns about bias. As the Task Force explores the use of these tools, we offer our perspective on how they can be used to address the risk of bias, which we hope will also inform the broader conversation at the hearing today.

As this Task Force is aware, the data-driven nature of AI makes it susceptible to unintentional bias. Because AI is trained on data from the past, there is a risk that AI systems may replicate and potentially further entrench historical inequities. As AI is integrated into business processes that can have consequential impacts on people's lives, there is a risk that "biased" systems will perform less accurately or unfairly disadvantage members of historically marginalized communities.

For BSA members, earning trust and confidence in the AI and other software they develop is crucial, so identifying ways to reduce the risk of bias is a priority. BSA therefore set out to develop real, credible, and actionable steps to guard against the potential of AI systems producing unintended disparate impacts. The resulting framework – Confronting Bias: BSA's Framework to Build Trust in AI – was released in June and is built on a vast body of research and informed by the experience of leading AI developers.[5]

The Framework outlines a lifecycle-based approach for performing impact assessments to identify risks of AI bias and corresponding best practices for mitigating those risks. The foundation of the Framework is its detailed methodology for performing impact assessments that help ensure that critical decisions are documented and that an organization's product development team, its compliance personnel, and senior leadership are aligned on the appropriate steps for mitigating risks of bias when they are identified. The Framework is intended to scale with risk and recognizes that inherently low-risk systems—for example, a system used to predict the type of fonts being used on a document—may not require a full impact assessment. But for systems that pose heightened risks, a robust impact assessment is essential to mitigating potential harms.

---

[3] Detecting Credit Card Fraud Using SMLE | Splunk; Splunk at TransUnion | Splunk

[4] See, e.g., Adobe - Adobe unveils new AI ethics principles as part of commitment to responsible digital citizenship; IBM - 3 lessons from IBM on designing ethical AI technology | World Economic Forum (weforum.org); Microsoft - Our approach to responsible AI at Microsoft; Salesforce - Salesforce Debuts AI Ethics Model: How Ethical Practices Further Responsible Artificial Intelligence - Salesforce News; Workday - Building Trust in AI and ML Through Principles, Practice, and Policy (workday.com).

[5] Confronting Bias: BSA's Framework to Build Trust in AI

The Framework is ultimately a playbook that organizations can use to enhance trust in their AI systems through risk management processes that promote fairness, transparency, and accountability, three of the key principles for responsible and ethical AI. The full Framework, with more than 50 actionable diagnostic statements, is attached to my testimony and can be found at ai.bsa.org. Below, I share a few key insights from the Framework.

Overview

AI is used in so many different contexts that only a flexible, risk management approach will be successful. The BSA Framework is built on three key elements:

(1) Identifying the risks of bias through **impact assessments** across a system's lifecycle;
(2) **Mitigating those risks** through concrete, actionable practices; and
(3) Setting forth key corporate governance structures to promote **organization accountability**.

Among the unique features of the BSA Framework is that it recognizes these elements need to be followed at all stages of the AI lifecycle: Design, Development, and Deployment and Use phases. Further, there are a variety of AI development and deployment models, and the Framework recognizes that the appropriate allocation of risk management responsibilities will vary depending on the type of system, including who develops the algorithm, trains the model, and ultimately deploys the system.

- ***AI Bias Can Arise Throughout the AI Lifecycle***

To combat AI bias, it is essential to understand the many sources of risk and the variety of ways they can manifest in an AI system. While much attention has understandably focused on data as a source of bias, the potential vectors of risk precede data collection efforts and begin at the earliest stages of a system's conception and design.

The initial step in building an AI system is often referred to as "problem formulation." It involves the identification and specification of the "problem" the system is intended to address, an initial mapping of how the model will achieve that objective, and the identification of a "target variable" the system will be used to predict. Because many AI systems are designed to make predictions about attributes that are not directly measurable, data scientists must often identify variables that can be used as proxies for the quality or outcome it is intended to predict.

While the use of proxy target variables can be entirely reasonable, the assumptions underlying the choice of proxies must be closely scrutinized to ensure that it does not introduce unintended bias to the system. The risk that can arise during this process of problem formulation is perhaps best exemplified by a recent study of a widely used healthcare algorithm that hospitals rely on to identify patients in need of urgent care. The research team concluded that the algorithm was systematically assigning lower risk scores to black patients compared to similarly sick white counterparts because it relied on data about historical healthcare costs as proxy for predicting a patient's future healthcare needs. Unfortunately, because black patients have historically had less access to healthcare, the reliance of spending data painted an inaccurate picture and led to dangerously biased outcomes.[6]

[6] Millions of black people affected by racial bias in health-care algorithms (nature.com)

The data used to train an AI system is a second major vector for bias. If the data used to train a system is misrepresentative of the population in which it will be used, there is a risk the system will perform less effectively on communities that may be underrepresented in the training data. Likewise, reliance on data that itself may be the product of institutional or historical biases can entrench those inequities in an AI model. The process of "labelling" training data can also introduce bias. Many AI systems require training data to be "labeled" so that the learning algorithm can identify patterns and correlations that can be used to classify future data inputs. Because the process of labeling the data can involve subjective decisions, there is the potential for introducing unintended bias into the training data.

Finally, even a system thoroughly vetted during development can begin to exhibit bias after it is deployed. AI systems are trained on data that represents a static moment in time and that filters out "noise" that could undermine the model's ability to make consistent and accurate predictions. Upon deployment in the real world, AI systems inevitably encounter conditions that differ from those in the development and testing environment. Further, because the real-world changes over time, the snapshot in time that a model represents may naturally become less accurate as the relationship between data variables evolves. If the input data for a deployed AI system differs materially from its training data, there is a risk that the system could "drift" and that the performance of the model could be undermined in ways that will exacerbate the risks of bias. For instance, if an AI system is designed (and tested) for use in a specific country, the system may not perform well if it is deployed in a country with radically different demographics. Bias can also arise if an AI system is deployed into an environment that differs significantly from the conditions for which it was designed or for purposes that are inconsistent with its intended use.

- ***Combatting AI Bias Requires a Lifecycle-Based Approach to Risk Management***

Although the challenges of AI bias are significant and without simple solutions, they are not insurmountable. Efforts to combat bias must start by recognizing that the issue requires a lifecycle-based approach to risk management.

Risk management is a process for ensuring systems are trustworthy by design by establishing a methodology for identifying risks and mitigating their potential impact. Risk management processes are particularly important in contexts, such as cybersecurity and privacy, where the combination of quickly evolving technologies and a highly dynamic threat landscapes render traditional "compliance" based approaches ineffective. Rather than evaluating a product or service against a static set of prescriptive requirements that quickly become outdated, risk management seeks to integrate compliance responsibilities into the development pipeline to help mitigate risks throughout a product or service's lifecycle.

But, what does that all mean in practice?

Companies that develop or use high-risk AI systems should establish a comprehensive approach for performing impact assessments. Impact assessments are widely used in a range of other fields—from environmental protection to data protection—as an accountability mechanism that promotes trust by demonstrating that a system has been designed in a manner that accounts for the potential risks it may pose to the public. The purpose of an impact assessment is to establish organizational processes to guide the development and use of high-risk systems by requiring internal stakeholders to identify the risks that a system may pose,

quantify the degree of harm the system could generate, and document any steps that have been taken to mitigate those risks to an acceptable level. By establishing a process for personnel to document key design choices and their underlying rationale, impact assessments are an important transparency and accountability mechanism.

The impact assessment methodology in the BSA Framework includes more than 40 diagnostic statements that should be documented throughout an AI system's lifecycle. Among its key recommendations is for organizations to maintain documentation about:

- The objectives and assumptions of the system, including its intended use cases and its target variable;
- The metrics that will be used as a baseline for evaluating bias in the system;
- The provenance of the data used to train the system, an evaluation of its appropriateness for the intended use case, and the steps that were taken to scrutinize the data for biases;
- The rationale for selecting data attributes and their impact on model performance; and
- The lines of responsibility for monitoring the system following deployment and plans for responding to potential incidents or system errors.

- ***Risk Management is a Collective Responsibility***

The documentation created and maintained as part of an impact assessment also facilitates important communication between the multiple stakeholders that may have roles to play managing AI risks. In many instances, the risk of bias may emerge at the intersection of system design decisions that were made by the system's developer and downstream decisions by the organizations that may deploy that system.

For instance, some AI developers provide general-purpose AI functionality, such as text analytics tools, that their customers can access and integrate into their own products and services via an API. In such a circumstance, risk management responsibilities will necessarily be shared by the system developer and the organization that deployed it. In other situations, the customers may, for privacy or other purposes, not allow the developer to view or assess data that may be used to re-train or fine tune the AI model.

While the precise allocation of risk management responsibilities will vary depending on the use case, as a general matter AI developers will be best positioned to provide information about the system's design and capabilities to enable the deployer to make informed deployment and risk mitigation decisions.

- ***Mitigating AI Bias Requires Diverse, Interdisciplinary Expertise***

A common refrain in the BSA Framework relates to the vital role of diversity in AI risk management efforts. Effectively identifying potential sources of bias in data requires a diverse set of expertise and experiences, including familiarity with the domain from which data is drawn and a deep understanding of the historical context and institutions that produced it. Moreover, oversight processes are most effective when team members bring diverse perspectives and backgrounds that can help anticipate the needs and concerns of users who may be impacted by or interact with an AI system.

Because "algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values," it is vital for organizations to establish teams that reflect a diversity of lived experiences and that traditionally underrepresented perspectives are included throughout the lifecycle of the AI design and development process.[7] To the extent an organization is lacking in diversity, it should consult with outside stakeholders to solicit feedback, particularly from underrepresented groups that may be impacted by the system.

Policy Recommendations

Public trust is an essential component of a thriving digital economy. While the responsibility for managing the risks of AI falls squarely on the organizations that develop and use AI systems, government can help foster public trust through policies that enhance the benefits of the technology while safeguarding against its potential risks. In the near term, we would advise Congress and the Administration to focus on the following lines of effort.

(1) Ensure consumer and civil rights protections remain fit-for-purpose in the digital age. Decisions that would otherwise be unlawful should not avoid liability simply because they may now involve the use of an AI system. To that end, we have encouraged efforts to audit federal agencies' existing consumer protection authorities to assess whether technological innovation is impeding their ability to enforce the law.[8] And we wrote to this Task Force last year about concerns that a rulemaking at HUD may exacerbate the risk of bias and discrimination in the housing market.[9]

(2) Establish a requirement for organizations to perform impact assessments prior to deploying high-risk AI systems. The BSA Framework can be one useful roadmap for new legislation.

(3) Promote international alignment around AI policy. Given the inherently global nature of the technology ecosystem, it is vital for the US to engage with our trading partners to forge consensus approaches for tackling shared challenges. There is an emerging global consensus that AI regulation should be risk-based and context specific. The EU recently introduced comprehensive legislation along these lines. The US should look for opportunities to drive these conversations, including through NIST's development of an AI risk management framework.

(4) Continue to emphasize privacy and security. Ethics and issues of bias are part of the trust formula, but privacy and data security laws are also essential.

---

[7] Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020): 33–44, https://doi.org/10.1145/3351095.3372873.

[8] *See* BSA Comments on Office of Management and Budget's Guidance on AI Regulation | BSA | The Software Alliance

[9] US: BSA Letter to the House Financial Services Committee Regarding Equitable Algorithms Hearing

Conclusion

Digital transformation across industry sectors is creating jobs and improving our lives. But industry, civil society, academia, and the government must work together on guidelines and laws that will ensure companies act responsibly in how they develop and deploy AI.

We appreciate the Task Force's strong focus on issues of ethics and bias. Confronting Bias: BSA's Framework to Build Trust in AI is our attempt to contribute meaningfully to this discussion. Thank you again for the opportunity to testify.

The Software Alliance

BSA

Confronting Bias: BSA's Framework to Build Trust in AI

www.bsa.org

CONTENTS

Introduction 1

What Is AI Bias? 3

Sources and Types of AI Bias 4

The Need for AI Risk Management 8

What Is Risk Management? 8

Managing the Risk of Bias 9

Foundations for Effective Risk Management 10

Governance Framework 11

Impact Assessment 13

AI Bias Risk Management Framework 14

AI Lifecycle Phases 15

Framework Structure 17

Stakeholder Roles and Responsibilities 18

Spectrum of AI Development and Deployment Models 18

BSA AI Bias Risk Management Framework 19

Foundational Resources 28

Endnotes 29

Introduction

Tremendous advances in artificial intelligence (AI) research and development are quickly transforming expectations about how the technology may shape the world. The promise that AI may one day impact every industry is quickly turning into a commercial reality. From financial services to healthcare, AI is increasingly leveraged to improve customer experiences, enhance competitiveness, and solve previously intractable problems. For instance, AI is enabling medical researchers to diagnose early-stage Alzheimer's Disease years before debilitating symptoms arise,[1] and it is helping ecologists analyze impossibly large datasets to better track the impact of their efforts to preserve critical habitat and prevent illegal elephant poaching in Malawi.[2]

As used in this report, the term "artificial intelligence" refers to systems that use machine learning algorithms that can analyze large volumes of training data to identify correlations, patterns, and other metadata that can be used to develop a model that can make predictions or recommendations based on future data inputs. For example, developers used machine learning to create "Seeing AI," an app that helps people who are blind or visually impaired navigate the world by providing auditory descriptions of objects in photographs.[3] Users of the app can use their smartphone to take pictures, and Seeing AI describes what appears in the photograph. To develop the computer vision model capable of identifying the objects in a picture, the system was trained using data from millions of publicly available images depicting common objects, such as trees, street signs, landscapes, and animals. When a user inputs a new image, Seeing AI in effect predicts what objects are in the photo by comparing it to the patterns and correlations that it derived from the training data.

The proliferation of AI across industries is also prompting questions about the design and use of the technology and what steps can be taken to ensure it is operating in a manner that accounts for any potential risks it may pose to the public.

The use of advanced technologies in connection with high-stakes decisions presents both opportunities and risks. On the one hand, the adoption of AI by financial institutions has the potential to reduce discrimination and promote fairness by facilitating a data-driven approach to decision-making that is less vulnerable to human biases.[4] For instance, the use of AI can improve access to credit and housing to historically marginalized communities by enabling lenders to evaluate a greater array of data than is ordinarily accounted for in traditional credit reports. At the same time, researchers caution that flaws in the design, development, and/or deployment of AI systems have the potential to perpetuate (or even exacerbate) existing societal biases.[5]

Developing mechanisms for identifying and mitigating the risks of AI bias has therefore emerged as an area of intense focus for experts in industry, academia, and government. In just the past few years, a vast body of research has identified a range of organizational best practices, governance safeguards, and technical tools that can help manage the risks of bias throughout the AI lifecycle. Static evaluations of AI models cannot account for all potential issues that may arise when AI systems are deployed in the field, so experts agree that mitigating risks of AI bias requires a lifecycle approach that includes ongoing monitoring by end-users to ensure that the system is operating as intended.

This document sets forth an *AI Bias Risk Management Framework* that organizations can use to perform impact assessments to identify and mitigate potential risks of bias that may emerge throughout an AI system's lifecycle. Similar to impact assessments for data privacy, AI impact assessments can serve as an important assurance mechanism that promotes accountability and enhances trust that high-risk AI systems have been designed, developed, tested, and deployed with sufficient protections in place to mitigate the risk of harm. AI impact assessments are also an important transparency mechanism that enables the many potential stakeholders involved in the design, development, and deployment of an AI system to communicate about its risks and ensure that responsibilities for mitigating those risks are clearly understood.

In addition to setting forth a process for performing an AI impact assessment, the Bias Risk Management Framework:

- Sets out the key corporate governance structures, processes, and safeguards that are needed to implement and support an effective AI risk management program; and
- Identifies existing best practices, technical tools, and resources that stakeholders can use to mitigate specific AI bias risks that can emerge throughout an AI system's lifecycle.

This Framework is intended to be a flexible tool that organizations can use to enhance trust in their AI systems through risk management processes that promote fairness, transparency, and accountability.

What Is AI Bias?

References to "AI bias" in this document refer to AI systems that systematically and unjustifiably yield less favorable, unfair, or harmful outcomes to members of specific demographic groups.

At its core, the goal of machine learning is to create a model that derives generalized rules from historical examples in order to make predictions about future data inputs. For instance, an image recognition system designed to identify plants would likely be trained on large volumes of photographs depicting each of the many species of vegetation. The system would look for general rules, like leaf patterns, that are common across the photographs of each species, thereby creating a model that can evaluate whether new data inputs (i.e., user-submitted photos) include any of the species it has been trained to identify. In other words, machine learning works by drawing generalizations from past data to make predictions about future data inputs. However, when AI is used to model human behavior, concerns about unintended bias take on an entirely different dimension. As AI is integrated into business processes that can have consequential impacts on people's lives, there is a risk that "biased" systems will systematically disadvantage members of historically marginalized communities. AI bias can manifest in systems that perform less accurately or treat people less favorably based on a sensitive characteristic, including but not limited to race, gender identity, sexual orientation, age, religion, or disability.

Sources and Types of AI Bias

DESIGN

AI bias can be introduced at multiple stages in the AI lifecycle.[6] Decisions made at the earliest stages of the conception and design of an AI system can introduce bias:

- **Problem Formulation Bias.** In some instances, the basic assumptions underlying a proposed AI system may be so inherently biased that they render it inappropriate for any form of public deployment.

EXAMPLES

In 2016, researchers at Shanghai Jiao Tong University published a highly controversial paper[7] detailing their effort to train an AI system to predict "criminality" through a facial imaging system. By training the system on a large volume of police mugshots, the researchers alleged that their system could predict "criminality" with close to 90 percent accuracy merely by analyzing a person's facial structure. Unsurprisingly, the paper quickly became the subject of scathing criticism, and commentators rightfully noted that the model relied on the profoundly disturbing (and causally unsupportable) assumption that criminality can be inferred from a person's appearance.[8]

Problem formulation bias can also arise when an AI system's target variable is an imprecise or overly simplistic proxy for what the system is actually trying to predict. For example, in 2019 researchers discovered that an AI system widely used by hospitals to triage patients[9] by predicting the likelihood that they required urgent care systematically prioritized the needs of healthier white patients to the detriment of less-healthy minority patients. In this instance, bias arose because the system sought to predict "healthcare needs" using historical data about "healthcare costs" as an easy-to-obtain stand-in for the actual data about the healthcare needs of patients. Unfortunately, because minority patients have historically had less access to healthcare, using "healthcare costs" as a proxy for the current needs of those patients paints an inaccurate picture that can result in dangerously biased outcomes.

- **Historical Bias.** There is a risk of perpetuating historical biases reflected in data used to train an AI system.

EXAMPLE

A medical school in the United Kingdom set out to create a system that would help identify good candidates for admission. The system was trained using data about previously admitted students. It was discovered, however, that the school's historical admissions decisions had systematically disfavored racial minorities and females whose credentials were otherwise equal to other applicants. By training the model using data reflecting historical biases, the medical school inadvertently created a system that replicated those same biased admission patterns.[10]

- **Sampling Bias.** If the data used to train a system is misrepresentative of the population in which it will be used, there is a risk that the system will perform less effectively on communities that may have been underrepresented in the training data. This commonly occurs when sufficient quantities of representative data are not readily available, or when data is selected or collected in ways that systematically over- or under-represent certain populations.

EXAMPLES

As the pathbreaking research by Joy Buolamwini and Timnit Gebru demonstrated, facial recognition systems trained on datasets composed disproportionately of white and male faces perform substantially less accurately when evaluating the faces of women with darker complexions.[11]

Sampling bias can also arise as a result of data collection practices. The City of Boston's attempt to create a system capable of automatically detecting and reporting potholes in need of repair is an illustrative case in point. Because early versions of the program relied heavily on data supplied by users of a smartphone app called "StreetBump," it received a disproportionate number of reports from affluent neighborhoods with residents who could afford smartphones and data plans. As a result of the sampling bias, potholes in poorer neighborhoods were underrepresented in the dataset, creating a risk that the system would allocate repair resources in a manner that would treat members of those communities unfairly.[12]

- **Labeling Bias.** Many AI systems require training data to be "labeled" so that the learning algorithm can identify patterns and correlations that can be used to classify future data inputs. The process of labeling the training dataset can involve subjective decisions that can be a vector for introducing human biases into the AI system.

EXAMPLE

ImageNet is a database of more than 14 million images that have been categorized and labeled to enable AI researchers to train vision recognition systems. Although ImageNet has been a critical tool for advancing the state of the art in AI object recognition, recent scholarship has shone a light on how the database's categorization and labeling system can create significant risks of bias when it is used to train systems involving images of people. In *Excavating AI*,[13] Kate Crawford and Trevor Paglen demonstrated that the categories and data labels associated with the images of people in ImageNet reflect a range of "gendered, racialized, ableist, and ageist" biases that could be propagated in any AI system that uses them as training data. For instance, an AI system trained on ImageNet data was more likely to classify images of Black subjects as "wrongdoers" or "offenders."[14]

DEVELOPMENT

Once the necessary data has been collected, the development team must clean, process, and normalize the data so that it can be used to train and validate a model. Developers must also select a machine learning approach, or adapt an off-the-shelf model, that is appropriate for the nature of the data they are using and the problem they are trying to solve. This may involve building many different models using different approaches and then choosing the most successful among them.[15] Usually, the development team must also make choices about data parameters to make the model functional. For instance, data reflecting a numerical score may be converted to a "yes" or "no" answer by assigning a threshold—for example, scores equal or greater to X may be re-designated as a "yes," and scores below that threshold designated "no." Biases that can emerge during the development stage include the following:

- **Proxy Bias.** The process of selecting the input variables (i.e., "features") that the model will weigh as it is being trained is another critical decision point that can introduce bias. Even when sensitive demographic data is excluded, bias may be introduced if the system relies on features that are closely correlated to those traits, called proxies.

EXAMPLE

Even the use of seemingly benign features can introduce proxy bias due to their correlation with sensitive attributes. Researchers have shown, for instance, that information about whether a person owns a Mac or PC laptop may be predictive of their likelihood to pay back a loan.[16] A financial institution might therefore seek to include such a variable when building an AI system to screen potential loan applicants. However, the inclusion of that feature also introduces a significant risk of proxy bias because Mac ownership correlates closely to race. As a result, its inclusion could result in a system that systematically disfavors applicants based on a feature that is closely correlated to race but that is unrelated to actual credit risk.

- **Aggregation Bias.** Using a "one-size-fits-all" model that overlooks key variables can result in system performance that is optimized only for the dominant sub-group. Aggregation bias can arise if the model fails to account for underlying differences between sub-groups that materially impact a system's accuracy rates. Rare phenomena may be lost in averages and aggregates. Worse, models of aggregated populations may correctly predict different or even opposite behavior to modes of sub-groups of the same population, a phenomenon known as Simpson's Paradox.

EXAMPLE

The risk of aggregation bias is particularly acute in healthcare settings where diagnosis and treatment must often account for the unique manner in which medical conditions may impact people across racial and ethnic lines. For instance, because the risk of complications posed by diabetes varies wildly across ethnicities, an AI system used to predict the risks associated with diabetes may underperform for certain patients unless it accounts for these differences.[17]

DEPLOYMENT, MONITORING, AND ITERATION

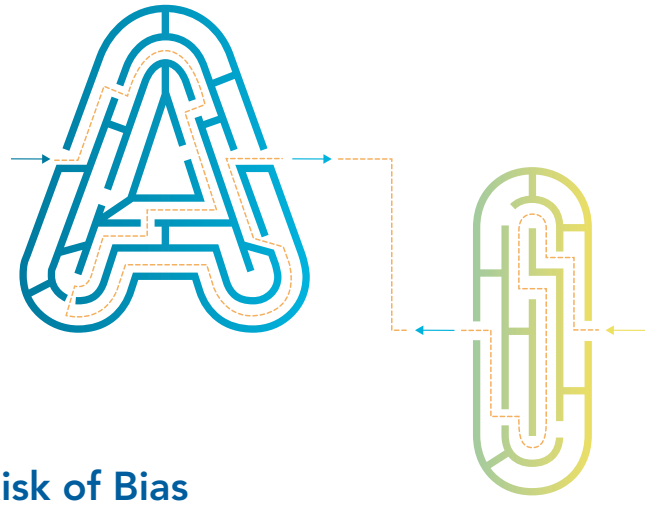
AI systems inevitably encounter real world scenarios that differ from the data used to train the model. As a result, even a system that has been thoroughly validated and tested prior to deployment may suffer performance degradation when it is put into production. Therefore, it is important that AI systems undergo ongoing evaluation and assessment throughout their lifecycles.

- **Deployment Bias.** Bias can arise in various ways after a system has been deployed, including when the data used to train or evaluate an AI system differs markedly from the population the system encounters when it is deployed, rendering the model unable to perform as intended. Deployment bias can emerge when a model is unable to reliably generalize beyond the data on which it was trained, either because the model was overfitted at the time of training (i.e., the prediction model learned so much detail about the training data that it is unable to make accurate generalizations about other data inputs) or because of concept drift (i.e., performance degradation was brought on by a shift in the relationship between the target variable and the training data).
- **Misuse Bias.** Deployment bias can also arise when an AI system or feature built for one purpose is used in an unexpected or unintended manner.

The Need for AI Risk Management

What Is Risk Management?

Risk management is a process for ensuring systems are trustworthy by design by establishing a methodology for identifying risks and mitigating their potential impact. Risk management processes are particularly important in contexts, such as cybersecurity and privacy, where the combination of quickly evolving technologies and highly dynamic threat landscapes render traditional "compliance" based approaches ineffective. Rather than evaluating a product or service against a static set of prescriptive requirements that quickly become outdated, risk management seeks to integrate compliance responsibilities into the development pipeline to help mitigate risks throughout a product or service's lifecycle. Effective risk management is anchored around a governance framework that promotes collaboration between an organization's development team and its compliance personnel at key points during the design, development, and deployment of a product.



Managing the Risk of Bias

Organizations that develop and use AI systems must take steps to prevent bias from manifesting in a manner that unjustifiably yields less favorable or harmful outcomes based on someone's demographic characteristics. Effectively guarding against the harms that might arise from such bias requires a risk management approach because:

"BIAS" AND "FAIRNESS" ARE CONTEXTUAL

It is impossible to eliminate bias from AI systems because there is no universally agreed upon method for evaluating whether a system is operating in a manner that is "fair." In fact, as Professor Arvind Narayanan has famously explained, there are at least 21 different definitions[18] (i.e., mathematical criteria) that can be used to evaluate whether a system is operating fairly, and it is *impossible* for an AI system to simultaneously satisfy all of them. Because no universal definition of fairness exists, developers must instead evaluate the nature of the system they are creating to determine which metric for evaluating bias is most appropriate for mitigating the risks that it might pose.

EFFORTS TO MITIGATE BIAS MAY INVOLVE TRADE-OFFS

Interventions to mitigate bias for one group can increase it for other groups and/or reduce a system's overall accuracy.[19] Risk management provides a mechanism for navigating such trade-offs in a context-appropriate manner.

BIAS CAN ARISE POST-DEPLOYMENT

Even if a system has been thoroughly evaluated prior to deployment, it may produce biased results if it is misused or deployed in a setting in which the demographic distribution differs from the composition of its training and testing data.

Foundations for Effective Risk Management

The aim of risk management is to establish repeatable processes for identifying and mitigating potential risks that can arise throughout an AI system's lifecycle. A comprehensive risk management program has two key elements:

1

A **governance framework** to support the organization's risk management functions.

2

A scalable process for performing an **impact assessment** to identify and mitigate risks.

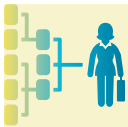
Governance Framework

Effective AI risk management should be underpinned by a governance framework that establishes the policies, processes, and personnel that will be used to identify, mitigate, and document risks throughout the system's lifecycle. The purpose of such a governance framework is to promote understanding across organizational units—including product development, compliance, marketing, sales, and senior management—about each entity's role and responsibilities for promoting effective risk management during the design, development, and deployment of AI systems. Key features of a risk management governance framework include:

Policies and Processes

At the core of the governance framework is a set of formal policies setting forth the organization's approach to risk management. These policies should define the organization's risk management objectives, the procedures that it will use to meet those objectives, and the benchmarks it will rely on for evaluating compliance.

- **Objectives.** AI risk management should be contextualized within an organization's broader risk management functions with the goal of ensuring that the organization is developing and using AI in a manner that aligns with its core values. To that end, the governance framework should identify how the organization will manage risks that could undermine those values.
- **Processes.** The governance framework should establish processes and procedures for identifying risks, assessing the materiality of those risks, and mitigating risks at each stage of the AI lifecycle.
- **Evaluation Mechanisms.** The governance framework should establish mechanisms, such as metrics and benchmarks, that the organization will use to evaluate whether policies and procedures are being carried out as specified.
- **Periodic Review.** As AI capabilities continue to mature and the technology is put to new uses, it is important that organizations periodically review and update their AI governance framework so that it remains fit-for-purpose and capable of addressing the evolving landscape of risk.



Executive Oversight. AI Developers and AI Deployers should maintain a governance framework that is backed by sufficient executive oversight. In addition to developing and approving the substance of the governance framework's policies, senior management should play an active role in overseeing the company's AI product development lifecycle. For high-risk systems that may negatively impact people in consequential ways, company leadership should be accountable for making "go/no-go" decisions.

Personnel, Roles, and Responsibilities

The effectiveness of risk management depends on establishing a cross-functional group of experts that can guide decisions throughout the AI lifecycle. Depending on the size of an organization and the nature of the systems it is developing or deploying, the responsibilities for risk management may involve staff from multiple business units. The governance framework should therefore identify the personnel within the organization who have roles and responsibilities related to AI risk management and clearly map reporting lines, authorities, and necessary expertise. In assigning roles and responsibilities, organizations should prioritize independence, competence, influence, and diversity.

- **Independence.** Risk management is most effective when personnel are structured in a manner that facilitates separate layers of independent review. For instance, risk management responsibilities may be split between multiple teams, including:
  - **Product Development Team.** Engineers, data scientists, and domain experts involved in designing and developing AI products and services.
  - **Compliance Team.** A diverse team of legal, compliance, domain experts, and data professionals who are responsible for overseeing compliance with the company's AI development policies and practices, such as the development of impact assessments for high-risk AI systems.
  - **Governance Team.** Ideally a senior management-led team with responsibility for developing, maintaining, and ensuring effective oversight of the organization's AI Governance Framework and risk management processes.
- **Competence, Resourcing, and Influence.** Personnel with risk management responsibilities must be provided with adequate training and resources to fulfill their governance functions. It is equally important to ensure that personnel are empowered and have the right incentives to make decisions to address and/or escalate risks. For instance, the organization should establish a clear escalation path that enables risk management personnel to engage with executive decision-makers so that there is executive-level visibility into key risk areas and decisions.

Diversity. The sociotechnical nature of AI systems makes it vitally important to prioritize diversity within the teams involved in a system's development and oversight. Development and oversight processes are most effective when team members bring diverse perspectives and backgrounds that can help anticipate the needs and concerns of users who may be impacted by or interact with an AI system. Because "algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values," it is vital that organizations establish teams that reflect a diversity of lived experiences and that traditionally underrepresented perspectives are included throughout the lifecycle of the AI design and development process.[20] To the extent an organization is lacking in diversity, it should consult with outside stakeholders to solicit feedback, particularly from underrepresented groups that may be impacted by the system.

Impact Assessment

To effectively manage AI risks, organizations should implement a robust process for performing impact assessments on any system that may materially impact members of the public. Impact assessments are widely used in a range of other fields—from environmental protection to data protection—as an accountability mechanism that promotes trust by demonstrating that a system has been designed in a manner that accounts for the potential risks it may pose to the public. In short, the purpose of an impact assessment is to identify the risks that a system may pose, quantify the degree of harm the system could generate, and document any steps that have been taken to mitigate those risks to an acceptable level.

Impact assessment processes should be tailored to address the nature of the system that is being evaluated and the type of harms it may pose. For truly low-risk systems—for example, a system used to predict the type of fonts being used on a document—a full impact assessment may not be necessary. But for systems that pose an inherent risk of material harm to the public, a full impact assessment should be performed. Given the incredible range of applications to which AI can be applied, there is no "one-size-fits-all" approach for identifying and mitigating risks. Instead, impact assessment processes should be tailored to address the nature of an AI system and the type of inherent risks and potential harms it may pose. To determine whether a system poses an inherent risk of material harm, stakeholders should consider:

- **Potential Impact on People.** Impact assessments are likewise important in circumstances where an AI system will be used in decision-making processes that may result in consequential impacts on people, such as their ability to obtain access to credit or housing.
- **Context and Purpose of the System.** Evaluating the nature of the AI system and the setting in which it will be used is a good starting point for determining both the necessity and appropriate scope of an impact assessment. Impact assessments are particularly critical for high-risk AI systems that will be used in domains (e.g., healthcare, transportation, finance) where the severity and/or likelihood of potential harms is high.
- **Degree of Human Oversight.** The degree to which an AI system is fully automated may also impact the inherent risks that it poses. A system designed to provide recommendations to a highly skilled professional is likely to pose fewer inherent risks than a similarly situated fully automated system. Of course, the mere existence of a human-in-the-loop certainly does not mean that an AI system is free from risk. It is necessary instead to examine the nature of the human-computer interaction holistically to determine the extent to which human oversight may mitigate an AI system's inherent risks.
- **Type of Data.** The nature of the data used to train a system can also shed light on a system's inherent risks. For instance, using training data relating to human characteristics or behaviors is a signal that a system may require closer scrutiny for bias.

AI Bias Risk Management Framework

We outline below an AI Bias Risk Management Framework that is intended to aid organizations in performing impact assessments on systems with potential risks of AI bias. In addition to setting forth processes for identifying the sources of bias that can arise throughout an AI system's lifecycle, the Framework identifies best practices that can be used to mitigate those risks.

The Framework is an assurance-based accountability mechanism that can be used by AI Developer and AI Deployer organizations for purposes of:

- **Internal Process Guidance.** AI Developers and AI Deployers can use the Framework as a tool for organizing and establishing roles, responsibilities, and expectations for internal processes.
- **Training, Awareness, and Education.** AI Developers and AI Deployers can use the Framework to build internal training and education programs for employees involved in developing and using AI systems. In addition, the Framework may provide a useful tool for educating executives about the organization's approach to managing AI bias risks.
- **Assurance and Accountability.** AI Developers and AI Deployers can use the Framework as a basis for communicating and coordinating about their respective roles and responsibilities for managing AI risks throughout a system's lifecycle.
- **Vendor Relations.** AI Deployers may choose to use the Framework to guide purchasing decisions and/or developing vendor contracts that ensure AI risks have been adequately accounted for.
- **Trust and Confidence.** AI Developers may wish to communicate information about a product's features and its approach to mitigating AI bias risks to a public audience. In that sense, the Framework can help organizations communicate to the public about their commitment to building ethical AI systems.
- **Incident Response.** Following an unexpected incident, the processes and documentation set forth in the Framework provide an audit trail that can help AI Developers and AI Deployers identify the potential source of system underperformance or failure.

AI Lifecycle Phases

The Framework is organized around the phases of the AI lifecycle, which represent the key iterative steps involved in the creation and use of an AI system.

DESIGN PHASE

- **Project Conception.** The initial stage of AI design involves identifying and formulating the "problem" that the system is intended to address and initially mapping how the model will achieve that objective. During this phase, the design team will define the purpose and structure of the system. Depending on the nature of the system, the design team will identify a target variable that the system is intended to predict. For instance, a fitness app that analyzes a consumer's heart rate to monitor for irregularities that might predict whether that person is at risk of a stroke or heart disease (i.e., the target variable). At this early stage of the system design process, the goal of the Bias Risk Management Framework is to identify whether using AI is appropriate for the project at hand. Potential risks include:
  - **Problem Formulation Bias.** Target variables may reflect inherent prejudices or faulty assumptions that can perpetuate harmful biases. In some instances, the basic assumptions underlying a proposed AI system may be so inherently biased as to render it inappropriate for any form of public deployment.
- **Data Acquisition.** Once the system objectives have been defined, developers must assemble a corpus of data that will be used to train the model to identify patterns that will enable it to make predictions about future data inputs. This training data can inadvertently introduce biases into an AI system in many ways. Potential risks include:
  - **Historical Bias.** Training an AI system using data that itself may reflect historical biases creates a risk of further entrenching those inequities.
  - **Sampling Bias.** The risk of bias also arises when the data used to train an AI system is not representative of the population in which it will be deployed. An AI system trained on unrepresentative data may not operate as effectively when making predictions about a member of a class that is either over- or under-represented.
  - **Labeling Bias.** Many AI systems require training data to be labeled so that it can identify what patterns it should be looking for. The process of labeling the training dataset can be a vector for introducing bias into the AI system.

DEVELOPMENT PHASE

- **Data Preparation and Model Definition.** The next step of the AI lifecycle involves preparing the data so that it is ready to train the model. During this process, the development team will clean, normalize, and identify the variables (i.e., "features") in the training data that the algorithm will evaluate as it looks for patterns and relationships as the basis of a rule for making future predictions. The team must also establish the system's underlying architecture, including selecting the type of algorithmic model that will power the system (e.g., linear regression, logistic regression, deep neural network.)[21] Once the data is ready and the algorithm is selected, the team will train the system to produce a functional model that can make predictions about future data inputs. Potential risks include the following:
  - **Proxy Bias.** The process of selecting features in the training data and choosing a modeling approach involves human decisions about what variables should be considered as relevant for making predictions about the model's target variable. These interventions can inadvertently introduce bias to the system, including by relying on variables that act as proxies for protected classes.
  - **Aggregation Bias.** Aggregation bias can arise if the model fails to account for underlying differences between sub-groups that materially impact a system's accuracy rates. Using a "one-size-fits-all" model that overlooks key variables can result in system performance that is optimized only for the dominant sub-group.
- **Model Validation, Testing, and Revision.** After the model has been trained, it must be validated to determine if it is operating as intended and tested to demonstrate that the system's outputs do not reflect unintended bias. Based on outcome of validation and testing, the model may need to be revised to mitigate risks of bias that are deemed unacceptable.

DEPLOYMENT PHASE

- **Deployment and Use.** Prior to deployment, the AI Developer should evaluate the system to determine whether risks identified in earlier stages of design and development have been sufficiently mitigated in a manner that corresponds to the company's governance policies. To the extent identified risks may arise through misuse of the system, the AI Developer should seek to control for them by integrating product features (e.g., user interfaces that reduce risk of misuse) to mitigate those risks, prohibiting uses that could exacerbate risks (e.g., end-user license agreements), and providing AI Deployers with sufficient documentation to perform their own impact assessments.

  Prior to using an AI system, an AI Deployer should review documentation provided by the AI Developer to assess whether the system corresponds with its own AI governance policies and to determine whether deployment-related risk management responsibilities are clearly assigned.

Although some post-deployment risk management responsibilities may be addressed by the AI Developer, the AI Deployer will often bear responsibility for monitoring system performance and evaluating whether it is operating in a manner that is consistent with its risk profile. Potential risks include:

- **Deployment Bias.** AI systems are trained on data that represents a static moment in time and that filters out "noise" that could undermine the model's ability to make consistent and accurate predictions. Upon deployment in the real world, AI systems will necessarily encounter conditions that differ from those in the development and testing environment. Further, because the real-world changes over time, the snapshot in time that a model represents may naturally become less accurate as the relationship between data variables evolves. If the input data for a deployed AI system differs materially from its training data, there is a risk that the system could "drift" and that the performance of the model could be undermined in ways that will exacerbate the risks of bias. For instance, if an AI system is designed (and tested) for use in a specific country, the system may not perform well if it is deployed in a country with radically different demographics.
- **Misuse Bias.** Deploying an AI system into an environment that differs significantly from the conditions for which it was designed or for purposes that are inconsistent with its intended use cases can exacerbate risks of bias.

Framework Structure

The Framework identifies best practices for identifying and mitigating risks of AI bias across the entire system lifecycle. It is organized into:

- **Functions,** which denote fundamental AI risk management activities at their highest level, dividing them between Impact Assessment and Risk Mitigation Best Practices.
- **Categories,** which set out the activities and processes that are needed to execute upon the Functions at each phase of the AI Lifecycle. In other words, the Categories set forth the steps for performing an Impact Assessment and identify the corresponding Risk Mitigation Best Practices that can be used to manage associated risks.
- **Diagnostic Statements,** which set forth the discrete actions that should be taken to execute upon the Categories. They provide a set of results that help support achievement of the outcomes in each Category.
- **Comments on Implementation,** which provide additional information for achieving the outcomes described in the Diagnostic Statements.
- **Tools and Resources,** which identify a range of external guidance and toolkits that stakeholders can use to mitigate the bias risks associated with each phase of the AI lifecycle. The specific tools and resources identified in the framework are non-exhaustive and are highlighted for informational purposes only.

Stakeholder Roles and Responsibilities

Reflecting the inherently dynamic nature of AI systems, the Framework is intended to account for the array of stakeholders that may play a role in various aspects of a system's design, development, and deployment. Because there is no single model of AI development or deployment, it is impossible in the abstract to assign roles or delegate specific responsibilities for many of the Framework's risk management functions. However, in general, there are three sets of stakeholders that may bear varying degrees of responsibility for certain aspects of AI risk management throughout a system's lifecycle:

- **AI Developers.** AI Developers are organizations responsible for the design and development of AI systems.
- **AI Deployers.** AI Deployers are the organizations that adopt and use AI systems. (If an entity develops its own system, it is both the AI Developer and the AI Deployer.)
- **AI End-Users.** AI End-Users are the individuals—oftentimes an employee of an AI Deployer—who are responsible for overseeing the use of an AI system.

The allocation of risk management responsibilities between these stakeholders will in many cases depend on an AI system's development and deployment model.

Spectrum of AI Development and Deployment Models

The appropriate allocation of risk management responsibilities between stakeholders will vary depending on the nature of the AI system being developed and which party determines the purposes and means by which the underlying model is trained. For instance:

- **Universal, Static Model.** The AI Developer provides all its customers (i.e., AI Deployers) with a static, pre-trained model.
  - The AI Developer will bear responsibility for most aspects of model risk management.
- **Customizable Model.** The AI Developer provides a pre-trained model to AI Deployers who can customize and/or retrain the model using their own data.
  - Risk management will be a shared responsibility between the AI Developer and the AI Deployer.
- **Bespoke Model.** The AI Developer trains a bespoke AI model on behalf of an AI Deployer using the AI Deployer's data.
  - Risk management will be a shared responsibility between the AI Developer and the AI Deployer, with the bulk of obligations falling on the AI Deployer.

BSA AI Bias Risk Management Framework

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
PROJECT CONCEPTION			
Impact Assessment	Identify and Document Objectives and Assumptions	Document the intent and purpose of the system.	• What is the purpose of the system—i.e., what "problem" will it solve? • Who is the intended user of the system? • Where and how will the system be used? • What are the potential misuses?
		Clearly define the model's intended effects.	What is the model intended to predict, classify, recommend, rank, or discover?
		Clearly define intended use cases and context in which the system will be deployed.	
	Select and Document Metrics for Evaluating Fairness	Identify "fairness" metrics that will be used as a baseline for assessing bias in the AI system.	The concept of "fairness" is highly subjective and there are dozens of metrics by which it can be evaluated. Because it is impossible to simultaneously satisfy all fairness metrics, it is necessary to select metrics that are most appropriate for the nature of the AI system that is being developed and consistent with any applicable legal requirements. It is important to document the rationale by which fairness metrics were selected and/or excluded to inform latter stages of the AI lifecycle.
	Document Stakeholder Impacts	Identify stakeholder groups that may be impacted by the system.	Stakeholder groups include AI Deployers, AI End-Users, Affected Individuals (i.e., members of the public who may interact with or be impacted by an AI system).
		For each stakeholder group, document the potential benefits and potential adverse impacts, considering both the intended uses and reasonably foreseeable misuses of the system.	
		Assess whether the nature of the system makes it prone to potential bias-related harms based on user demographics.	User demographics may include, but are not limited to race, gender, age, disability status, and their intersections.
	Document Risk Mitigations	If risk of bias is present, document efforts to mitigate risks.	

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
PROJECT CONCEPTION			
Impact Assessment *(continued)*	Document Risk Mitigations	Document how identified risks and potential harms of each risk will be measured and how the effectiveness of mitigation strategies will be evaluated.	
		If risk of bias is present, document efforts to mitigate risks.	
		If risks are unmitigated, document why the risk was deemed acceptable.	
Risk Mitigation Best Practices	Independence and Diversity	Seek feedback from a diverse set of stakeholders to inform the impact assessment.	Because risks identified during this initial phase will inform later aspects of the development and impact assessment processes, it is vital to develop a holistic understanding of potential harms that may arise by soliciting diverse perspectives from people with a range of lived experiences, cultural backgrounds, and subject matter expertise. To the extent in-house personnel lack subject matter or cultural diversity, it may be necessary to consult with third-party experts or to solicit feedback from members of communities that may be adversely impacted by the system.
	Transparent Documentation	Share impact assessment documentation with personnel working on later stages of the AI pipeline so that risks and potential unintended impacts can be monitored throughout the development process.	
	Accountability and Governance	Ensure that senior leadership has been adquately briefed on potential high risk AI systems.	Impact assessment documentation for systems deemed "high risk" should be shared with senior leadership to facilitate a "go/no-go" decision.
DATA ACQUISITION			
Impact Assessment	Maintain Records of Data Provenance	Maintain sufficient records to enable "recreation" of the data used to train the AI model, verify that its results are reproducible, and monitor for material updates to data sources.	Records should include: • Source of data • Origin of data (e.g., Who created it? When? For what purpose? How was it created?) • Intended uses and/or restrictions of the data and data governance rules (e.g., What entity owns the data? How long can it be retained (or must it be destroyed)? Are there restrictions on its use?) • Known limitations of data (e.g., missing elements?) • If data is sampled, what was the sampling strategy? • Will the data be updated? If so, will any versions be tracked?

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
DATA ACQUISITION			
Impact Assessment *(continued)*	Examine Data for Potential Biases	Scrutinize data for historical biases.	Examine sources of data and assess potential that they may reflect historical biases.
		Evaluate "representativeness" of the data.	• Compare demographic distribution of training data to the population where the system will be deployed. • Assess whether there is sufficient representation of subpopulations that are likely to interact with the system.
		Scrutinize data labeling methodology.	• Document personnel and processes used to label data. • For third-party data, scrutinize labeling (and associated methodologies) for potential sources of bias.
	Document Risk Mitigations	Document whether and how data was augmented, manipulated, or re-balanced to mitigate bias.	
Risk Mitigation Best Practices	Independence and Diversity	To facilitate robust interrogation of the datasets, data review teams should include personnel that are diverse in terms of their subject matter expertise and lived experiences.	Effectively identifying potential sources of bias in data requires a diverse set of expertise and experiences, including familiarity with the domain from which data is drawn and a deep understanding of the historical context and institutions that produced it. To the extent in-house personnel lack diversity, consultation with third-party experts or potentially affected stakeholder groups may be necessary.
	Re-Balancing Unrepresentative Data	Consider re-balancing with additional data.	Improving representativeness can be achieved in some circumstances by collecting additional data that improves the balance of the overall training dataset.
		Consider re-balancing with synthetic data.	Imbalanced datasets can potentially be rebalanced by "oversampling" data from the underrepresented groups. A common oversampling method is the Synthetic Minority Oversampling Technique, which generates new "synthesized" data from the underrepresented group.

DESIGN

Function	Category	Diagnostic Statement	Comments on Implementation
DATA ACQUISITION			
Risk Mitigation Best Practices *(continued)*	Data Labeling	Establish objective and scalable labeling guidelines.	• To mitigate the potential of labeling bias, the personnel responsible for labeling the data should be provided with clear guidelines establishing an objective and repeatable process for individual labeling decisions. • In domains where the risk of bias is high, labelers should have adequate subject matter expertise and be provided training to recognize potential unconscious biases. • For high-risk systems, it may be necessary to set up a quality assurance mechanism to monitor label quality.
	Accountability and Governance	Integrate data labeling processes into a comprehensive data strategy.	Establishing an organizational data strategy can help ensure that data evaluation is performed consistently and prevent duplication of effort by ensuring that company efforts to scrutinize data are documented for future reference.

DESIGN: RISK MITIGATION TOOLS AND RESOURCES

Project Conception

- ***Aequitas Bias and Fairness Audit Toolkit***
  Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy (2018), http://www.datasciencepublicpolicy.org/projects/aequitas/.
- ***Diverse Voices Project | A How-To Guide for Facilitating Inclusiveness in Tech Policy***
  Lassana Magassa, Meg Young, and Batya Friedman, University of Washington Tech Policy Lab, https://techpolicylab.uw.edu/project/diverse-voices/.

Data Compilation

- ***Datasheets for Datasets***
  Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, arXiv:1803.09010v7, (March 19, 2020), https://arxiv.org/abs/1803.09010.
- ***AI FactSheets 360***
  IBM Research, https://aif360.mybluemix.net/.

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
DATA PREPARATION AND MODEL DEFINITION			
Impact Assessment	Document Feature Selection and Engineering Processes	Document rationale for choices made during the feature selection and engineering processes and evaluate their impact on model performance.	Examine whether feature selection or engineering choices may rely on implicitly biased assumptions.
		Document potential correlation between selected features and sensitive demographic attributes.	For features that closely correlate to a sensitive class, document the relevance to the target variable and the rationale for its inclusion in the model.
	Document Model Selection Process	Document rationale for the selected modeling approach.	
		Identify, document, and justify assumptions in the selected approach and potential resulting limitations.	
Risk Mitigation Best Practices	Feature Selection	Examine for biased proxy features.	• Simply avoiding the use of sensitive attributes as inputs to the system—an approach known as "fairness through unawareness"—is not an effective approach to mitigating the risk of bias. Even when sensitive characteristics are explicitly excluded from a model, other variables can act as proxies for those characteristics and introduce bias into the system. To avoid the risk of proxy bias, the AI Developer should examine the potential correlation between a model's features and protected traits and examine what role these proxy variables may be playing in the model's output. • The ability to examine statistical correlation between features and sensitive attributes may be constrained in circumstances where an AI Developer lacks access to sensitive attribute data and/or is prohibited from making inferences about such data.[22] In such circumstances, a more holistic analysis informed by domain experts may be necessary.

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
DATA PREPARATION AND MODEL DEFINITION			
Risk Mitigation Best Practices *(continued)*	Feature Selection	Scrutinize features that correlate to sensitive attributes.	• Features that are known to correlate to a sensitive attribute should only be used if there is a strong logical relationship to the system's target variable. • For example, income—although correlated to gender—is reasonably related to a person's ability to pay back a loan. The use of income in an AI system designed to evaluate creditworthiness would therefore be justified. In contrast, the use of "shoe size"—which also correlates to gender—in a model for predicting creditworthiness would be an inappropriate use of a variable that closely correlates to a sensitive characteristic.
	Independence and Diversity	Seek feedback from diverse stakeholders with domain-specific expertise.	The feature engineering process should be informed by personnel with diverse lived experiences and expertise about the historical, legal, and social dimensions of the data being used to train the system.
	Model Selection	Avoid inscrutable models in circumstances where both the risk and potential impact of bias are high.	Using more interpretable models can mitigate the risks of unintended bias by making it easier to identify and mitigate problems.
VALIDATING, TESTING, AND REVISING THE MODEL			
Impact Assessment	Document Validation Processes	Document how the system (and individual components) will be validated to evaluate whether it is performing consistent with the design objectives and intended deployment scenarios.	
		Document re-validation processes.	• Establish cadence at which model will be regularly re-validated. • Establish performance benchmarks that will trigger out-of-cycle re-validation.
	Document Testing Processes	Test the system for bias by evaluating and documenting model performance.	Testing should incorporate fairness metrics identified during Design phase and examine the model's accuracy and error rates across demographic groups.
		Document how testing was performed, which fairness metrics were evaluated, and why those measures were selected.	
		Document model interventions.	If testing reveals unacceptable levels of bias, document efforts to refine the model.

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
VALIDATING, TESTING, AND REVISING THE MODEL			
Risk Mitigation Best Practices	Model Interventions	Evaluate potential model refinements to address bias surfaced during testing.	In circumstances where testing reveals that the system is exhibiting unacceptable levels of bias based on the selected fairness metric, it will be necessary to refine the model. Potential model refinements include: • **Pre-Processing Interventions.** Such refinements can involve revisiting earlier stages of the Design and Development lifecycle (e.g., seeking out additional training data). • **In-Processing Interventions.** Bias can also be mitigated by imposing an additional fairness constraint directly on the model. Traditional machine learning models are designed to maximize for predictive accuracy. Emerging techniques enable developers to build constraints into the model to reduce the potential for bias across groups. The addition of a fairness constraint, in effect, instructs the model to optimize both for accuracy and a specific fairness metric. • **Post-Processing Interventions.** In some cases, bias can be addressed through the use of post-processing algorithms that manipulate the model's output predictions to ensure that it adheres to a desired distribution.
	Independence and Diversity	Validation and testing documentation should be reviewed by personnel who were not involved in the system's development.	The independent team should compare the validation and testing results to the system specifications developed during earlier phases of the design and development process.

DEVELOPMENT: RISK MITIGATION TOOLS AND RESOURCES

- ***Model Cards for Model Reporting***
  Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, (January 2019): 220–229, https://arxiv.org/abs/1810.03993.
- ***AI Factsheets 360***
  Aleksandra Mojsilovic, IBM Research (August 22, 2018), https://www.ibm.com/blogs/research/2018/08/factsheets-ai/.
- ***AI Explainability 360***
  IBM Research, https://aix360.mybluemix.net/.
- ***AI Fairness 360***
  IBM Research, https://aif360.mybluemix.net/.
- ***Responsible Machine Learning with Error Analysis***
  Besmira Nushi, Microsoft Research (February 18, 2021), https://techcommunity.microsoft.com/t5/azure-ai/responsible-machine-learning-with-error-analysis/ba-p/2141774.
- ***Aequitas Open Source Bias Audit Toolkit***
  Pedro Saleiro, Abby Stevens, Ari Anisfeld, and Rayid Ghani, University of Chicago Center for Data Science and Public Policy, http://www.datasciencepublicpolicy.org/projects/aequitas/.
- ***FairTest: Discovering Unwarranted Associations in Data-Driven Applications***
  Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels and Huang Lin, ArXiv, (2015), https://github.com/columbia/fairtest.
- ***Bayesian Improved Surname Geocoding***
  Consumer Finance Protection Bureau (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

DEPLOYMENT AND USE

Function	Category	Diagnostic Statement	Comments on Implementation
PREPARING FOR DEPLOYMENT AND USE			
Impact Assessment	Document Lines of Responsibility	Define and document who is responsible for the system's outputs and the outcomes they may lead to, including details about how a system's decisions can be reviewed if necessary.	
		Establish management plans for responding to potential incidents or reports of system errors.	• What does it mean for the system to fail and who might be harmed by a failure? • How will failures be detected? • Who will respond to failures when they are detected? • Can the system be safely disabled? • Are there appropriate plans for continuity of critical functions?
	Document Processes for Monitoring Data	Document what processes and metrics will be used to evaluate whether production data (i.e., input data the system encounters during deployment) differs materially from training data.	
	Document Processes for Monitoring Model Performance	For static models, document how performance levels and classes of error will be monitored over time and benchmarks that will trigger review.	
		For models that are intended to evolve over time, document how changes will be inventoried; if, when, and how versions will be captured and managed; and how performance levels will be monitored (e.g., cadence of scheduled reviews, performance indicators that may trigger out-of-cycle review).	
	Document Audit and End-of-Life Processes	Document the cadence at which impact assessment evaluations will be audited to evaluate whether risk mitigation controls remain fit for purpose.	
		Document expected timeline that system support will be provided and processes for decommissioning system in event that it falls below reasonable performance thresholds.	
Risk Mitigation Best Practices	Monitoring for Drift and Model Degradation	Input data encountered during deployment can be evaluated against a statistical representation of the system's training data to evaluate the potential for data drift (i.e., material differences between the training data and deployment data that can degrade model performance).	

DEPLOYMENT AND USE

Function	Category	Diagnostic Statement	Comments on Implementation
PREPARING FOR DEPLOYMENT AND USE			
Risk Mitigation Best Practices *(continued)*	Product Features and User Interface	Integrate product and user interface features to mitigate risk of foreseeable unintended uses—e.g., interface that enforces human-in-the-loop requirements, alerts to notify when a system is being misused.	
	System Documentation	AI Developers should provide sufficient documentation regarding system capabilities, specifications, limitations, and intended uses to enable AI Deployers to perform independent impact assessment concerning deployment risks.	If necessary, AI Developers can also provide AI Deployers with a technical environment to perform an independent impact assessment.
		Consider incorporating terms into the End-User License Agreement that set forth limitations designed to prevent foreseeable misuses (e.g., contractual obligations to ensure end-user will comply with acceptable use policy).	
		Sales and marketing materials should be closely reviewed to ensure that they are consistent with the system's actual capabilities.	
	AI User Training	AI Deployers should provide training for AI Users regarding a system's capabilities and limitations, and how outputs should be evaluated and integrated into a workflow.	For human-in-the-loop oversight of AI system to be an effective risk mitigation measure, AI Users should be provided adequate information and training so they can understand how the system is operating and make sense of the model's outputs.
	Incident Response and Feedback Mechanisms	AI Deployers should maintain a feedback mechanism to enable AI Users and Affected Individuals (i.e., members of the public that may interact with the system) to report concerns about the operation of a system.	For consequential decisions, Affected Individuals should be provided with an appeal mechanism.

DEPLOYMENT AND USE: RISK MITIGATION TOOLS AND RESOURCES

- ***AI Incident Response Checklist***
  BNH.AI, https://www.bnh.ai/public-resources.
- ***Watson OpenScale***
  IBM, https://www.ibm.com/cloud/watson-openscale.
- ***Detect Data Drift on Datasets***
  Microsoft Azure Machine Learning (June 25, 2020), https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets?tabs=python#create-dataset-monitors.

Foundational Resources

A Framework for Understanding Unintended Consequences of Machine Learning
Harini Suresh and John V. Guttag, arXiv (February 2020), https://arxiv.org/abs/1901.10002.

AI Fairness
Trisha Mahoney, Kush R. Varshney, and Michael Hind, O'Reilly (April 2020), https://www.oreilly.com/library/view/ai-fairness/9781492077664/.

Beyond Explainability: A Practical Guide to Managing Risk in Machine Learning Models
Andrew Burt, Brenda Leong, Stuart Shirrell, and Xiangnong (George) Wang, Future of Privacy Forum (June 2018), https://fpf.org/wp-content/uploads/2018/06/Beyond-Explainability.pdf.

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI
Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach, CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (April 2020): 1–14, https://doi.org/10.1145/3313831.3376445.

Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing
Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P., FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (January 2020): 33–44, https://doi.org/10.1145/3351095.3372873.

Supervisory Guidance on Model Risk Management
US Federal Reserve Board (April 2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector
David Leslie, The Alan Turing Institute (2019), https://doi.org/10.5281/zenodo.3240529.

ENDNOTES

[1] Gina Kolata, "Alzheimer's Prediction May Be Found in Writing Tests," *New York Times* (February 1, 2021), https://www.nytimes.com/2021/02/01/health/alzheimers-prediction-speech.html.

[2] Dina Temple-Raston, *Elephants under Attack Have an Unlikely Ally: Artificial Intelligence*, NPR (October 25, 2019), https://www.npr.org/2019/10/25/760487476/elephants-under-attack-have-an-unlikely-ally-artificial-intelligence.

[3] *Seeing AI: An App for Visually Impaired People That Narrates the World Around You*, Microsoft, https://www.microsoft.com/en-us/garage/wall-of-fame/seeing-ai/.

[4] See e.g., Jennifer Sukis, *The Origins of Bias and How AI May Be the Answer to Ending Its Reign*, Medium (January 13, 2019), https://medium.com/design-ibm/the-origins-of-bias-and-how-ai-might-be-our-answer-to-ending-it-acc3610d6354.

[5] See e.g., Nicol Turner Lee, Paul Resnick, and Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, Brookings (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[6] Harini Suresh and John V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning* (February 17, 2020), https://arxiv.org/pdf/1901.10002.pdf.

[7] See Xiaolin Wu and Xi Zhang, *Automated Inference on Criminality Using Face Images*, Shanghai Jiao Tong University (November 13, 2016), https://arxiv.org/pdf/1611.04135v1.pdf.

[8] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov, *Physiognomy's New Clothes*, Medium (May 6, 2017), https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a.

[9] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* (October 25, 2019), https://science.sciencemag.org/content/366/6464/447.

[10] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California University Law Review* 104, no. 3 (September 30, 2016): 671, http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.

[11] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[12] Kate Crawford, *The Hidden Biases in Big Data*, Harvard Business Review (April 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data.

[13] Kate Crawford and Trevor Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets* (September 19, 2019), https://excavating.ai/.

[14] Cade Metz, "'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You?" *New York Times* (September 20, 2019), https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html.

[15] Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael Littman, *Model Selection's Disparate Impact in Real-World Deep Learning Applications*, arXiv:2104.00606 (April 1, 2021), https://arxiv.org/abs/2104.00606.

[16] Aaron Klein, *Credit Denial in the Age of AI*, Brookings Institution (April 11, 2019), https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/.

[17] J. Vaughn, A. Baral, M. Vadari "Analyzing the Dangers of Dataset Bias in Diagnostic AI systems: Setting Guidelines for Dataset Collection and Usage," ACM Conference on Health, Inference and Learning, 2020 Workshop, http://juliev42.github.io/files/CHIL_paper_bias.pdf.

[18] Arvind Narayanan, *21 Fairness Definitions and Their Politics*, ACM Conference on Fairness, Accountability and Transparency (March 1, 2018), https://www.youtube.com/watch?v=jIXluYdnyyk.

[19] Reuben Binns and Valeria Gallo, *AI Blog: Trade-Offs*, UK Information Commission's Office (July 25, 2019), https://ico.org.uk/about-the-ico/news-and-events/ai-blog-trade-offs/.

[20] Inioluwa Deborah Raji et al., *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (January 2020): 33–44, https://doi.org/10.1145/3351095.3372873.

[21] Sara Hooker, Moving Beyond "Algorithmic Bias Is a Data Problem," *Patterns* (April 9, 2021), https://www.sciencedirect.com/science/article/pii/S2666389921000611.

[22] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, *"What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness*, arXiv:2011.02282 (January 23, 2021), https://arxiv.org/abs/2011.02282.

www.bsa.org

BSA Worldwide Headquarters

20 F Street, NW
Suite 800
Washington, DC 20001

+1.202.872.5500
@BSAnews
@BSATheSoftwareAlliance

BSA Asia-Pacific

300 Beach Road
#30-06 The Concourse
Singapore 199555

+65.6292.2072

BSA Europe, Middle East & Africa

44 Avenue des Arts
Brussels 1040
Belgium

+32.2.274.13.10