



26 July 2023

BSA COMMENTS ON SUPPORTING SAFE AND RESPONSIBLE ARTIFICIAL INTELLIGENCE IN AUSTRALIA

Submitted Electronically to the Department of Industry, Science and Resources

BSA | The Software Alliance (**BSA**)¹ welcomes the opportunity to submit comments to the Department of Industry, Science and Resources (**DISR**) on its Discussion Paper pertaining to Supporting Safe and Responsible AI in Australia (**Discussion Paper**).²

BSA is the leading advocate for the global software industry. BSA members create technology solutions that power other businesses, including cloud storage services, customer relationship management software, human resources management programs, identity management services, security solutions, and collaboration software. Our members are on the leading edge of providing AI-enabled products and services, and tools used by others in the development of AI systems and applications. As a result, they have unique insights into the technology's tremendous potential to spur digital transformation and the policies that can best support the responsible use of AI.

We welcome DISR's efforts to explore how the Australian Government can support safe and responsible AI practices. As the Discussion Paper notes, AI presents significant opportunities and has delivered substantial benefits across both the economy and society, but also like any groundbreaking innovation, creates new risks and challenges.³ BSA's views on how to best mitigate the risks of AI are informed by our recent experience working with member companies to develop the BSA Framework to Build Trust in AI (**BSA Framework**),⁴ a risk management framework for mitigating the potential for unintended bias throughout an AI system's lifecycle. Built on a vast body of research and informed by the experience of leading AI developers and deployers, the BSA Framework outlines a lifecycle-based approach for performing impact assessments to identify risks of AI bias and highlights corresponding risk mitigation best practices. BSA has testified before the United States Congress and the European Parliament on the Framework and its approach to mitigating AI-related risks.⁵ Our extensive experience on these issues informs our response to the Discussion Paper.

¹ BSA's members include: Adobe, Alteryx, Altium, Amazon Web Services, Atlassian, Autodesk, Bentley Systems, Box, Cisco, Cloudflare, CNC/Mastercam, Dassault, Databricks, DocuSign, Dropbox, Elastic, Graphisoft, IBM, Informatica, Juniper Networks, Kyndryl, MathWorks, Microsoft, Nikon, Okta, Oracle, Prokon, PTC, Rockwell, Rubrik, Salesforce, SAP, ServiceNow, Shopify Inc., Siemens Industry Software Inc., Splunk, Trend Micro, Trimble Solutions Corporation, TriNet, Twilio, Unity Technologies, Inc., Workday, Zendesk, and Zoom Video Communications, Inc.

² Safe and Responsible AI in Australia Discussion Paper, June 2023, https://storage.googleapis.com/converlens-au-industry/p/prj2452c8e24d7a400c72429/public_assets/Safe-and-responsible-AI-in-Australia-discussion-paper.pdf

³ Discussion Paper (2023), p. 3 and p. 7.

⁴ Confronting Bias: BSA's Framework to Build Trust in AI, June 2021, <https://ai.bsa.org/wp-content/uploads/2021/06/2021bsaaibias.pdf>, and enclosed.

⁵ Testimony of Victoria Espinel, Public Hearing on AI & Bias, Special Committee on Artificial Intelligence in a Digital Age, European Parliament, November 2021, https://www.europarl.europa.eu/cmsdata/244265/AIDA_Verbatim_30_November_2021_EN.pdf; Testimony of Aaron Cooper, Task Force on Artificial Intelligence: Beyond I, Robot: Ethics, Artificial Intelligence and the Digital Age, before the House

Summary of BSA's Recommendations

BSA offers the following recommendations, which respond to the specific policy issues highlighted in the Discussion Paper.

Definitions

1. DISR should stay abreast of international efforts to define key terms, with a view to promoting harmonisation and interoperability.
2. AI policies should define, and consequently distinguish, AI developers and AI deployers. Both types of entities should have their respective obligations to ensure responsible AI innovation, and those obligations should be tailored to their different roles in the ecosystem.

Risk-based approaches

3. Obligations should fall on the entity that is best positioned to both identify and efficiently mitigate the risk of harm.
4. Regulatory efforts should be focused on high-risk AI use cases.

Strengthen risk management mechanisms

5. Impact assessments for high-risk uses of AI systems should play a significant role in AI risk management. BSA supports the development of AI policies that acknowledge the utility and incorporate the use of AI impact assessments.
6. DISR should draw on existing risk management frameworks, such as the US National Institute of Standards and Technology's (**NIST**) AI Risk Management Framework (**RMF**) and the BSA Framework and promote risk management tools to help Australian companies implement processes for managing risks of AI systems.

Target areas

7. To reduce the risk of AI bias, impact assessments should be conducted throughout the AI lifecycle.
8. Obligations aimed at enhancing transparency in AI systems should account for the different roles that AI developers and AI deployers play in the ecosystem.
9. Responsible AI innovation must be built on a solid foundation of data privacy. In the context of the proposed obligations on AI and ADM in the Privacy Act Review, BSA recommends: a) introducing a clear definition for "legal or similarly significant" that is interoperable with rights created in privacy laws internationally; and b) ensuring that these rights are exercised through the data controller and not the data processor.

Definitions

1. Stay abreast of international efforts to define key terms

The Discussion Paper defines the following terms: AI, Machine Learning, Generative AI models, Large Language Models (**LLMs**), Multimodal Foundation Model (**MfM**) and Automated Decision Making (**ADM**).

BSA is encouraged that, in defining these terms, DISR has referred to definitions found in other internationally recognised standards and frameworks. Specifically, the Discussion Paper's definitions of AI and Machine Learning are based on the International Standards Organization's (**ISO**) definitions.⁶ In particular, the ISO definition of AI shares many similarities with the Organization for Economic Co-operation and Development's (**OECD**) definition of AI in their Recommendation of

Financial Services Committee, October 2021, <https://www.congress.gov/117/meeting/house/114125/witnesses/HHRG-117-BA00-Wstate-CooperA-20211013.pdf>.

⁶ ISO/IEC 22989: 2022 – Artificial Intelligence Concepts and Terminology, July 2022, <https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en>.

Council on Artificial Intelligence (**Recommendation**). The NIST AI RMF adapts the OECD definition and defines AI as an “engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments” and that operates “with varying levels of autonomy”.⁷ We support this definition.

Given that AI systems are developed and deployed in an international context, definitions pertaining to AI should ideally be aligned across jurisdictions to ensure that all stakeholders have a common understanding of AI. As part of the work of the U.S.-EU Trade and Technology Council, the United States and the EU are agreeing on shared interpretations of key defined terms. For example, the EU-U.S. Terminology and Taxonomy for Artificial Intelligence defines machine learning as a branch of AI “and computer science which focuses on development of systems that are able to learn and adapt without following explicit instructions imitating the way that humans learn, gradually improving its accuracy, by using algorithms and statistical models to analyse and draw inferences from patterns in data.”⁸

Prioritising international alignment in defining AI-related terms will: a) reduce discrepancies and conflicts between different legal frameworks, thus promoting compliance; b) serve as foundation for dialogue and cooperation between governments on AI-related risks; and c) support the international development of best practices and benchmarks for using AI systems safely, allowing AI systems to be deployed responsibly on a global scale.

Recommendation: DISR should stay abreast of international efforts to define key terms, with a view to promoting harmonisation and interoperability.

2. Define and distinguish “AI developers” and “AI deployers”

The Discussion Paper’s definitions did not account for the different entities involved in an AI system’s supply chain. Reflecting the inherently dynamic nature of AI systems, policies pertaining to AI must account for the array of stakeholders that may play a role in various aspects of a system’s design, development, and deployment. In general, there are at least two key stakeholders with varying degrees of responsibility for managing the risks associated with an AI system throughout its lifecycle:

- **AI developers:** An AI developer is an entity that designs, codes, or produces an AI system.
- **AI deployers:** An AI deployer is an entity that uses an AI system.

It is crucial to define, and consequently distinguish, the AI developer and the AI deployer. As explained in the OECD’s Recommendation, effective AI policies must necessarily account for “stakeholders according to their role and the context” in which AI is being deployed.⁹ Effective management of risks among these different actors will depend on the nature of the AI system being developed. Distinguishing between AI developers and AI deployers ensures that specified obligations reflect an entity’s role in the AI ecosystem. Tailoring obligations to an entity’s role as an AI developer or AI deployer enables the company to fulfill the corresponding obligations and better protect consumers.

For example, an AI developer that designs an AI system is well-positioned to have access to information about the type of data that is used to train an AI system, the system’s known limitations, and its intended use cases. However, the AI developer would *not* have insight into how the AI system is used after another company has purchased and deployed the AI system. Instead, the AI deployer – the entity using the AI system – is generally best positioned to provide details on how the system is

⁷ NIST Artificial Intelligence Risk Management Framework 1.0, January 2023, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>.

⁸ EU-U.S. Terminology and Taxonomy for Artificial Intelligence, May 2023, <https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence>.

⁹ OECD Recommendation (2019). Per the Recommendation, the AI stakeholder community “encompasses all organizations and individuals involved in, or affected by, AI systems, directly or indirectly.”

being used, the outputs from the AI system, the nature of any customer complaints, and other real-world factors affecting the system's performance. AI deployers are also best positioned to understand the risk profile that an AI system may present to individuals. Ensuring AI policies create obligations that reflect these different roles will enable all stakeholders to better understand how their organisations can identify and address harmful bias in AI systems.

Recommendation: AI policies should define, and consequently distinguish, AI developers and AI deployers. Both types of entities should have their respective obligations to ensure responsible AI innovation, and those obligations should be tailored to their different roles in the ecosystem.¹⁰

Risk-based approaches

3. Obligations should fall on appropriate entities

Obligations associated with AI should not apply to all actors in the AI ecosystem in the same way. BSA has emphasised the importance of distinguishing the AI developer from the AI deployer because, by implementing this distinction, it will minimise uncertainty about which actor will bear responsibility for complying with key legal requirements reflected in AI legislation or policies and better enable organisations to fulfill responsibilities according to their different roles.

Indeed, the regulatory requirements for the various entities responsible for developing and deploying AI should account for their unique roles and capabilities. Any obligation should fall on the entity that is best positioned to both identify and efficiently mitigate the risk of harm. We support AI governance approaches that promote accountability for both AI developers and deployers respectively, as each entity in the design, development and deployment of AI should have clear responsibilities.

For example, when using impact assessments to manage AI risks, both AI developers and deployers should document key aspects of AI systems, which are important reference points for understanding the operation of AI systems. However, the information to be documented will be different for developers that design an AI system than for deployers using an AI system:

- **Developers of high-risk AI systems** should document information including, as appropriate:
 - The intended purpose of the AI system;
 - Known limitations of the AI system;
 - Known, likely, and specific high risks that could occur and steps taken to mitigate those risks;
 - An overview of the data used to train the AI system; and
 - A summary of how the AI system was evaluated prior to sale.
- **Deployers of high-risk AI systems** should document information including, as appropriate:
 - The purpose for which the deployer intends to use the AI system;
 - Transparency measures, including notices to impacted individuals about the AI system's use;
 - A summary of how the AI system is evaluated, if applicable;
 - Known, likely, and specific high risks that could occur and steps taken to mitigate those risks; and
 - Post-deployment monitoring and user safeguards, if applicable.

Recommendation: Obligations should fall on the entity that is best positioned to both identify and efficiently mitigate the risk of harm.

¹⁰ For further information, please see: AI Developers and Deployers: An Important Distinction, March 2023, <https://www.bsa.org/files/policy-filings/03162023aidevdep.pdf>

4. Focus on high-risk AI use cases

The AI ecosystem is broad, encompassing a diverse range of technologies and use cases and a wide array of stakeholders. Because the risks of AI are inherently use-case specific, any regulations should focus on specific applications of the technology that pose higher risks to the public but should be flexible enough to account for the unique considerations that may be implicated by specific use cases.

As a general principle, the scope of any regulatory obligations should be a function of the degree of risk and the potential scope and severity of harm. Many AI systems pose extremely low, or even no, risk to individuals or society, while creating potentially significant benefits. Imposing onerous regulations on such low-risk systems would hamper AI innovation with few corresponding benefits and therefore limit opportunities to use AI for positive impact. For example, AI is a critical component of cybersecurity risk mitigation,¹¹ which creates significant benefits to both companies and to consumers. Regulation should be mindful of the unintended consequences of regulations that could inadvertently limit the deployment of AI in detecting and responding to ever-evolving cybersecurity threats.

AI regulatory efforts should focus on addressing high-risk AI use cases. In this regard, AI systems may be high risk if they are used to make decisions to hire, promote, or terminate an individual's employment, or in other contexts, to determine eligibility for credit, healthcare, or housing.

Recommendation: Regulatory efforts should be focused on high-risk AI use cases.

Strengthen risk management mechanisms

5. Impact assessments should play a significant role in AI risk management

The Discussion Paper identified a range of AI-related risks, including concerns regarding algorithmic bias, misleading outputs, privacy, and transparency.¹² In view of these risks, accountability mechanisms, which enable organisations to demonstrate that they have developed or deployed AI responsibly, are crucial to maintain public confidence in AI systems. While the Discussion Paper highlighted various regulatory options,¹³ BSA notes that the use of impact assessments was only briefly considered as part of a draft risk-based approach for AI.¹⁴

Impact assessments should play a significant role in DISR's approach to AI risk management. Impact assessments are an important accountability mechanism that exists today and can be applied to AI, as they can be used to help AI developers and deployers of AI systems for high-risk uses identify and mitigate risks throughout the lifecycle of an AI system. By allowing personnel across the organisation to examine the objectives, data preparation, design choices, and testing results, impact assessments help to drive internal changes to an organisation's risk management program. Implementing these changes enables organisations to better address existing concerns and adapt to new risks as they emerge. The fact that assessments are being performed for high-risk uses of AI systems also promotes trust for external stakeholders because they will know that an organisation is conducting a thorough examination of AI systems, and that the assessments are available to regulators upon request in the event of an investigation.

A recent report on AI accountability also concluded that impact assessments had several advantages over other accountability tools, noting that: 1) they are familiar to organisations already conducting impact assessments for privacy and data protection; 2) they are practical because they do not rely on technical standards, which are currently nascent; and 3) they are future-proof because they can adapt

¹¹ An organisation could face millions of indicators of compromise per day and security teams demand contextual awareness and visibility from across their entire environments. Cybersecurity providers that leverage AI can detect and respond to both known and unknown threats in real-time, with speed and scale to match.

¹² Discussion Paper (2023), p. 7-8.

¹³ Discussion Paper (2023), p. 28 – 31.

¹⁴ Discussion Paper (2023), p. 40.

as AI systems and AI governance evolve.¹⁵

Notably, the Discussion Paper indicates that final results of an impact assessment should be published and that external experts should conduct reviews for high-risk uses. This approach creates concerns, because one reason that impact assessments create a strong accountability tool is that they are conducted internally through confidential assessments. One of the goals of impact assessments is to drive internal changes, and organisations will likely conduct less thorough reviews to surface problems if the results ultimately will either be made public or be shared with third parties. In addition, impact assessments will likely contain confidential business information that organisations want to protect. Further, sharing sensitive information with third parties could implicate privacy and security concerns. Accordingly, we recommend that the process of conducting an impact assessment remain an internal exercise without the involvement of third parties or publication of the results.

For the reasons discussed above, BSA strongly supports the use of impact assessments to mitigate risks arising from high-risk uses of AI systems. The BSA Framework similarly recommends the use of impact assessments for high-risk uses of AI systems and is appropriate for organisations to implement to enhance AI accountability.

Recommendation: Impact assessments should play a significant role in AI risk management for high-risk use cases. BSA supports the development of AI policies that acknowledge the utility and incorporate the use of AI impact assessments for high-risk uses.

6. Promote and use risk management tools

BSA supports a risk-based approach for addressing potential AI risks. Risk management frameworks and tools developed in collaboration with stakeholders across the AI ecosystem can help drive consensus around processes for identifying, measuring, mitigating, and communicating about AI risks. DISR should draw on existing risk management frameworks, such as the NIST AI RMF and the BSA Framework, and promote risk management tools to help Australian companies implement processes for managing risks of AI systems.

In this regard, the BSA Framework identifies specific practices that AI developers and deployers can implement to ensure that AI is developed and used responsibly. For example, the BSA Framework highlights a range of issues that impact assessments should address, including identifying fairness metrics that will be used to assess bias in the AI system, ensuring that senior leadership has been briefed on potential high-risk AI systems, scrutinising training data for bias, and documenting how testing was performed.

BSA has also strongly supported NIST's work to develop its AI RMF, in line with our longstanding recognition that risk management is a key component of promoting trust in AI. The NIST AI RMF identifies specific practices that can be implemented to develop and deploy trustworthy AI. Both frameworks recommend:

- Consultation with a diverse group of stakeholders;
- Establishing processes to identify, assess, and mitigate risks;
- Defining individual roles and responsibilities to people throughout an organisation;
- Identifying metrics for evaluation;
- Evaluating fairness and bias;
- Maintaining post-deployment feedback mechanisms; and
- Establishing detailed plans for responding to incidents.

BSA recently prepared a crosswalk between the two frameworks, which illustrates the significant

¹⁵ Impact Assessments: Supporting AI Accountability & Trust, January 2023, <https://accesspartnership.com/impact-assessments-supporting-ai-accountability/>

alignment between the two approaches.¹⁶ These frameworks are appropriately voluntary, and the flexibility to utilise different components allows organisations to tailor the frameworks to their business needs.

Recommendation: DISR should draw on existing risk management frameworks, such as the NIST AI RMF and the BSA Framework, and promote risk management tools to help Australian companies implement processes for managing risks of AI systems.

Target Areas

7. Reducing the risk of bias in AI systems

As highlighted in the Discussion Paper, “[a]lgorithmic bias is often raised as one of the biggest risks or dangers of AI,” involving “systematic or repeated decisions that privilege one group over another.”¹⁷

The BSA Framework elaborates on how bias could occur at multiple stages in the development and deployment of an AI system.¹⁸ For example, with respect to training data, there are risks of perpetuating historical biases reflected in the data or of sampling bias – where the data is misrepresentative of the population in which it will be used. 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. Bias can also arise in various ways after a system has been deployed, including when the data used to train or evaluate an AI system differs materially from the population the system encounters when it is deployed.

One way to identify and address these risks in high-risk systems is conducting an impact assessment. For example, the BSA Framework recommends that in the data acquisition phase of the design of an AI model, companies should evaluate the representativeness of the data as part of conducting an impact assessment. To do so, a company can compare the demographic distribution of training data to the population in which the system will be deployed and assess whether there is sufficient representation of subpopulations that are likely to interact with the system. To mitigate issues that arise, companies can consider rebalancing the dataset with additional data or synthetic data, which involves oversampling data from underrepresented groups.

Similarly, in the data preparation and model definition phase, an impact assessment could include documenting a potential correlation between selected features and sensitive demographic attributes. For features that closely correlate to a sensitive class, companies can document the relevance to the target variable and the rationale for its inclusion in the model, consistent with laws prohibiting discriminatory actions. These practices, along with other measures, can help to prevent algorithmic discrimination while enabling innovation.

Recommendation: To reduce the risk of AI bias, impact assessments should be conducted to address issues throughout the AI lifecycle.

8. Enhancing transparency in AI systems

The Discussion Paper notes the importance of transparency in ensuring “appropriate accountability, risk mitigation and responsibility for liability is applied appropriately across AI vendors and buyers along the value chain.”¹⁹

BSA supports enhancing transparency of AI systems. As policy approaches to enhance transparency

¹⁶ Crosswalk Between BSA Framework to Build Trust in AI and NIST AI Risk Management Framework, April 2023, <https://www.bsa.org/policy-filings/us-crosswalk-between-bsa-framework-to-build-trust-in-ai-and-nist-ai-risk-management-framework>

¹⁷ Discussion Paper (2023), p. 8.

¹⁸ BSA Framework (2021), p. 4-7.

¹⁹ Discussion Paper (2023), p. 8-9.

are developed, both AI developers and AI deployers play distinct but important roles in ensuring appropriate transparency for consumers. Notably, while AI developers are best positioned to understand the intended uses and limitations of an AI system, AI deployers are the entities that typically interact with consumers and decide when and how to deploy the AI system for a particular use. Their respective obligations should reflect this difference in roles – for example, an AI developer might be required to maintain documentation summarising how its AI system was evaluated prior to sale, whereas an AI deployer might be required to maintain documentation on post-deployment monitoring and user safeguards.

In regard to transparency, a developer might be obligated to make available to a deployer information reasonably necessary for the deployer to conduct an impact assessment, including information about the AI system's capabilities, limitations, and intended purpose. A deployer's transparency obligations might focus on the information provided to consumers about the use of an AI system.

Recommendation: Obligations aimed at enhancing transparency in AI systems should account for the different roles that AI developers and AI deployers play in the ecosystem.

9. Safeguarding privacy in AI systems

The data-intensive nature of AI makes privacy protections vital. The Discussion Paper observed that, while “[r]ich, large and quality data sets are a fundamental input to AI”, “access to and application of these datasets have the potential for individuals’ data to be used in ways that raise privacy concerns.”²⁰ Indeed, responsible AI innovation must be built on a solid foundation of data governance that prioritises the security of user data and aligns with the public’s expectation about how their personal information will be used. To engender trust in AI systems, consumers must have confidence in how their data is used and protected, and consumers’ expectations should be backstopped by strong legal obligations on companies that collect and/or process sensitive data.

Relatedly, BSA provided comments to the Attorney-General’s Department’s Privacy Act Review Report 2022 (**AGD Report**), where we addressed the AGD Report’s proposals to introduce several new obligations related to AI and automated decision making (**ADM**) that have “legal or similarly significant effect” on an individual’s rights.²¹ BSA supports introducing robust consumer rights and obligations on companies that handle data. However, we also highlighted the importance of: a) creating a comprehensive definition for the phrase “legal or similarly significant effects” that is interoperable with other privacy laws, so as to increase certainty for both individuals and companies when the related rights are available and allow them to better apply these protections across different jurisdictions; and b) individuals should be exercising these rights through a data controller, rather than a processor acting on behalf of a controller.²²

Recommendation: Responsible AI innovation must be built on a solid foundation of data privacy. In the context of the proposed obligations on AI and ADM in the Privacy Act Review, BSA recommends: a) introducing a clear definition for “legal or similarly significant” that is interoperable with rights created in privacy laws internationally; and b) ensuring that these rights are exercised through the data controller and not the data processor.

Conclusion

We hope that our comments will assist DISR in its objective to ensure AI is used safely and responsibly. We look forward to serving as a resource as you continue to engage in policy discussions on this issue. Please do not hesitate to contact me if you have any questions regarding this

²⁰ Discussion Paper (2023), p. 8.

²¹ BSA Comments on the Privacy Act Review Report 2022, April 2023, <https://www.bsa.org/files/policy-filings/04102023bsaupriv.pdf>, p. 10-11.

²² Data controllers are entities that decide how and why to collect information, which should be distinguished from data processors, which are entities that process collected personal information on behalf of other entities. The controller-processor distinction is similar to the AI developer-deployer distinction in that it distinguishes the consumer-facing entity (the AI deployer and the data controller) from the non-consumer facing entity (the AI developer and the data processor).

submission or if I can be of further assistance.

Sincerely,



Tham Shen Hong
Manager, Policy – APAC



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

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 paper⁷ 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 patients⁹ 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

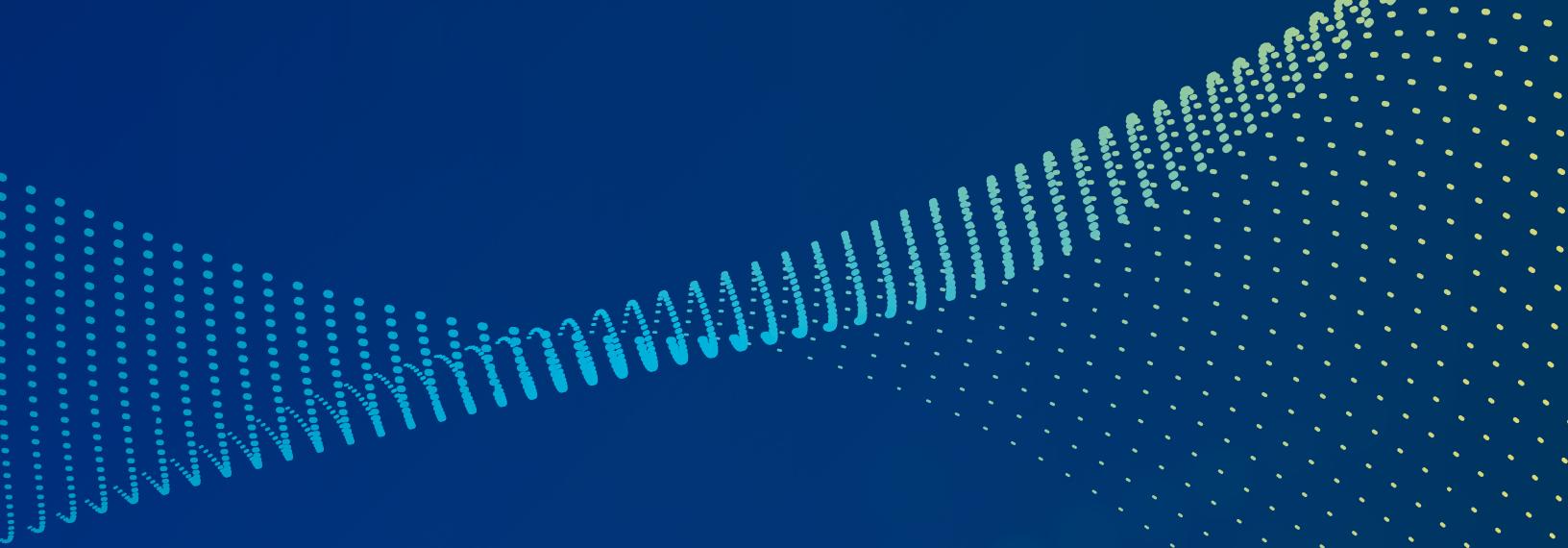
The risk of aggregation bias is particularly acute in healthcare settings where diagnosis and treatment must often account for the unique manner in which medical conditions may impact people across racial and ethnic lines. For instance, because the risk of complications posed by diabetes varies wildly 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.

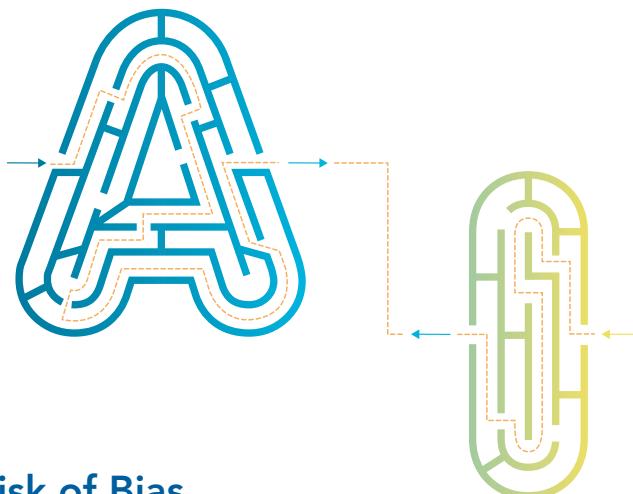
- **Deployment Bias.** Bias can arise in various ways after a system has been deployed, including when the data used to train or evaluate an AI system differs markedly from the population the system encounters when it is deployed, rendering the model unable to perform as intended. Deployment bias can 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 definitions¹⁸ (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	<p>Document the intent and purpose of the system.</p>	<ul style="list-style-type: none"> • 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?
		<p>Clearly define the model’s intended effects.</p>	What is the model intended to predict, classify, recommend, rank, or discover?
		<p>Clearly define intended use cases and context in which the system will be deployed.</p>	
	Select and Document Metrics for Evaluating Fairness	<p>Identify “fairness” metrics that will be used as a baseline for assessing bias in the AI system.</p>	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	<p>Identify stakeholder groups that may be impacted by the system.</p>	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).
		<p>For each stakeholder group, document the potential benefits and potential adverse impacts, considering both the intended uses and reasonably foreseeable misuses of the system.</p>	
		<p>Assess whether the nature of the system makes it prone to potential bias-related harms based on user demographics.</p>	User demographics may include, but are not limited to race, gender, age, disability status, and their intersections.
	Document Risk Mitigations	<p>If risk of bias is present, document efforts to mitigate risks.</p>	

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 adequately 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.	<p>Records should include:</p> <ul style="list-style-type: none"> • 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 <i>(continued)</i>	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.	<ul style="list-style-type: none"> 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.	<ul style="list-style-type: none"> 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.	<ul style="list-style-type: none"> 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.	<ul style="list-style-type: none"> 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 <i>(continued)</i>	Feature Selection	Scrutinize features that correlate to sensitive attributes.	<ul style="list-style-type: none"> 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.	<ul style="list-style-type: none"> 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.	<p>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:</p> <ul style="list-style-type: none"> 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.datasciencelpublicpolicy.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.	<ul style="list-style-type: none"> • 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.	
	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.bnhi.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,' 'Non-smoker,' '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 Kwagyir-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 Inioliwa 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