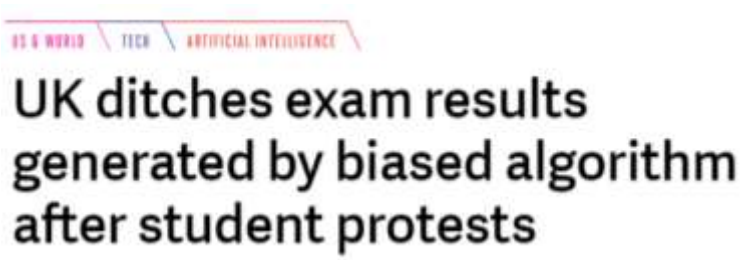

The Software Alliance
BSA
Confronting Bias:
BSA's Framework to Build Trust in AI
www.bsa.org

The Need to Address Bias

- Growth of AI across sectors
- Acceleration of digital transformation during the pandemic
- Increase in the number of people directly impacted by AI systems



What Is AI Bias?

Defining and Identifying Bias

- An AI system is "biased" if it
  - ***Systematically and unjustifiably yields less favorable, unfair, or harmful outcomes to members of specific demographic groups.***

- Manifestations of AI Bias
  - 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 of Bias
  - AI Bias can be introduced at multiple stages in the AI lifecycle, including **Design, Development** and **Deployment.**

AI Lifecycle – Potential Sources of Bias

Design

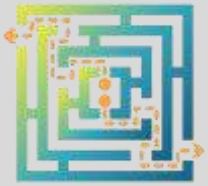
- Project Conception
- Data Acquisition

- *Problem Formulation Bias*
- *Historical Bias*
- *Sampling Bias*
- *Labeling Bias*

Development

- Data Preparation and Model Definition
- Validating, Testing, and Revising the Model

- *Proxy Bias*
- *Aggregation Bias*



Deployment and Use

- *Deployment Bias*
- *Misuse Bias*

Risk Management

- A process for ensuring systems are trustworthy by design by establishing a methodology for identifying risk and mitigating their potential impact.

The risk of AI bias must be managed because it is impossible to eliminate.

- "Bias" and "fairness" are contextual
- Efforts to mitigate bias may involve trade-offs
- Bias can arise post-deployment

Objectives

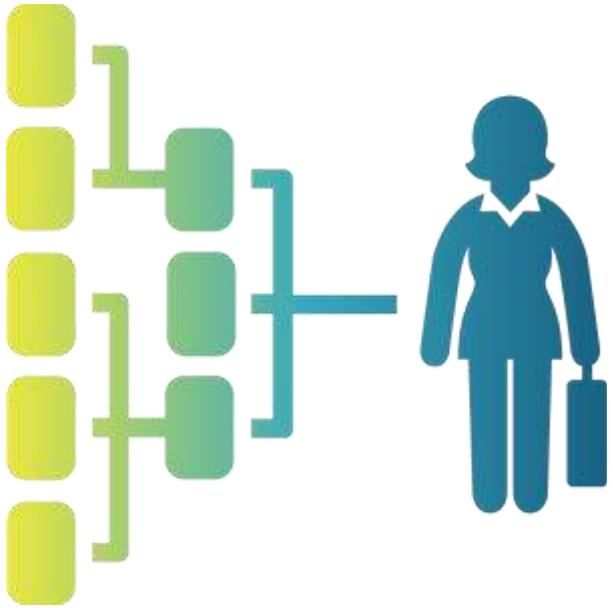
Objectives

- **Organizational Accountability – Governance Safeguards**
  - Key corporate governance structures, processes and safeguards for supporting an AI risk management program
- **Identifying Bias – Impact Assessment**
  - Process for performing AI impact assessments to identify the sources and potential risks of bias
- **Mitigating Bias – Best Practices**
  - Best practices, technical tools, and resources for mitigating AI bias risks

The Framework is tool for enhancing trust in AI systems through risk management processes that promote fairness, trust, and accountability.

Governance Framework

- **Policies and Processes**
  - Objectives
  - Processes
  - Evaluation Mechanisms
  - Periodic Review
  - Executive Oversight
- **Personnel, Roles and Responsibilities**
  - Independence
  - Competence, Resourcing, and Influence
  - Diversity



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.
  - *Potential Impact on People*
  - *Context and Purpose of the System*
  - *Degree of Human Oversight*
  - *Type of Data*

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.

Framework Excerpt

DESIGN

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

Framework Excerpt

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
DATA PREPARATION AND MODEL DEFINITION			
Impact Assessment	Document Feature Selection and Engineering Processes	Document rationale for choices made during the feature selection and engineering processes and evaluate their impact on model performance.	Examine whether feature selection or engineering choices may rely on implicitly biased assumptions.
		Document potential correlation between selected features and sensitive demographic attributes.	For features that closely correlate to a sensitive class, document the relevance to the target variable and the rationale for its inclusion in the model.
	Document Model Selection Process	Document rationale for the selected modeling approach.	
		Identify, document, and justify assumptions in the selected approach and potential resulting limitations.	
Risk Mitigation Best Practices	Feature Selection	Examine for biased proxy features.	• Simply avoiding the use of sensitive attributes as inputs to the system—an approach known as "fairness through unawareness"—is not an effective approach to mitigating the risk of bias. Even when sensitive characteristics are explicitly excluded from a model, other variables can act as proxies for those characteristics and introduce bias into the system. To avoid the risk of proxy bias, the AI Developer should examine the potential correlation between a model's features and protected traits and examine what role these proxy variables may be playing in the model's output. • The ability to examine statistical correlation between features and sensitive attributes may be constrained in circumstances where an AI Developer lacks access to sensitive attribute data and/or is prohibited from making inferences about such data.[1] In such circumstances, a more holistic analysis informed by domain experts may be necessary.
	Feature Selection	Scrutinize features that correlate to sensitive attributes.	• Features that are known to correlate to a sensitive attribute should only be used if there is a strong logical relationship to the system's target variable. • For example, income—although correlated to gender—is reasonably related to a person's ability to pay back a loan. The use of income in an AI system designed to evaluate creditworthiness would therefore be justified. In contrast, the use of "shoe size"—which also correlates to gender—in a model for predicting creditworthiness would be an inappropriate use of a variable that closely correlates to a sensitive characteristic.

Framework Excerpt

DEPLOYMENT AND USE

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

Policy Considerations + Challenges

Crafting laws and regulations to promote responsible AI…

- **Focusing on Risk**
  - Defining Systems and Use Cases
- **Assigning Responsibilities**
  - Multiple Stakeholders
  - Multiple Development Models
  - Range of Underlying Technologies and Use Cases
- **Avoiding Perverse Incentives**