AI Bias Risk Management Framework

DESIGN

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

DESIGN

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

DESIGN

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

DESIGN

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

DESIGN: RISK MITIGATION TOOLS AND RESOURCES

Project Conception

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

Data Compilation

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

DEVELOPMENT

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

[1] 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.

DEVELOPMENT

Function	Category	Diagnostic Statement	Comments on Implementation
DATA PREPARATION AND MODEL DEFINITION			
Risk Mitigation Best Practices *(continued)*	Independence and Diversity	Seek feedback from diverse stakeholders with domain-specific expertise.	The feature engineering process should be informed by personnel with diverse lived experiences and expertise about the historical, legal, and social dimensions of the data being used to train the system.
	Model Selection	Avoid inscrutable models in circumstances where both the risk and potential impact of bias are high.	Using more interpretable models can mitigate the risks of unintended bias by making it easier to identify and mitigate problems.
VALIDATING, TESTING, AND REVISING THE MODEL			
Impact Assessment	Document Validation Processes	Document how the system (and individual components) will be validated to evaluate whether it is performing consistent with the design objectives and intended deployment scenarios.	
		Document re-validation processes.	• Establish cadence at which model will be regularly re-validated. • Establish performance benchmarks that will trigger out-of-cycle re-validation.
	Document Testing Processes	Test the system for bias by evaluating and documenting model performance.	Testing should incorporate fairness metrics identified during Design phase and examine the model's accuracy and error rates across demographic groups.
		Document how testing was performed, which fairness metrics were evaluated, and why those measures were selected.	
		Document model interventions.	If testing reveals unacceptable levels of bias, document efforts to refine the model.

DEVELOPMENT

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

DEVELOPMENT: RISK MITIGATION TOOLS AND RESOURCES

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

DEPLOYMENT AND USE

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

DEPLOYMENT AND USE

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

DEPLOYMENT AND USE: RISK MITIGATION TOOLS AND RESOURCES

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