PROJECT DESCRIPTION

# MITIGATING AI/ML BIAS IN CONTEXT

## Establishing Practices for Testing, Evaluation, Verification, and Validation of AI Systems

Apostol Vassilev
Harold Booth
Murugiah Souppaya

National Institute of Standards and Technology

November 2022

ai-bias@nist.gov

This revision incorporates comments from the public.

The National Cybersecurity Center of Excellence (NCCoE), a part of the National Institute of Standards and Technology (NIST), is a collaborative hub where industry organizations, government agencies, and academic institutions work together to address businesses' most pressing cybersecurity and artificial intelligence (AI) challenges. Through this collaboration, the NCCoE develops modular, adaptable example cybersecurity and AI solutions demonstrating how to apply standards and best practices by using commercially available technology. To learn more about the NCCoE, visit https://www.nccoe.nist.gov/. To learn more about NIST, visit https://www.nist.gov/. To learn more about the NIST AI program, visit https://www.nist.gov/artificial-intelligence.

This document describes a challenge that is relevant across various industry sectors where machine learning (ML) assisted decision-making by humans is used, e.g., credit underwriting, employment hiring, school admissions. NIST AI experts and NCCoE experts will address this challenge through collaboration with members of the corresponding sectors, academic and industry researchers, civil society organizations, and technology providers of AI/ML-based decision-making systems that are currently deployed. The resulting reference design and guidance will detail an approach for mitigating bias that can be used by practitioners who deploy AI/ML technology for automating decision-making within organizations.

## ABSTRACT

Managing bias in an AI system is critical to establishing and maintaining trust in its operation. Despite its importance, bias in AI systems remains endemic across many application domains and can lead to harmful impacts regardless of intent. Bias is also context-dependent. To tackle this complex problem, we adopt a socio-technical approach to testing, evaluation, verification, and validation (TEVV) of AI systems in context. This approach connects the technology to societal values in order to develop recommended guidance for deploying AI/ML-based decision-making applications in a sector of the industry. This project will also look at the interplay between bias and cybersecurity. The project will leverage existing commercial and open-source technology in conjunction with the NIST Dioptra, an experimentation test platform for ML datasets and models. The initial phase of the project will focus on a proof-of-concept implementation for credit underwriting decisions in the financial services sector. This project will result in a freely available NIST AI/ML Practice Guide.

## KEYWORDS

*AI-assisted human decision-making; AI bias; AI fairness; artificial intelligence (AI); bias detection; bias mitigation; credit underwriting; human-computer interaction; machine learning (ML); machine learning model*

## DISCLAIMER

Certain commercial entities, equipment, products, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST or NCCoE, nor is it intended to imply that the entities, equipment, products, or materials are necessarily the best available for the purpose.

## TABLE OF CONTENTS

# 1 Executive Summary

## Purpose

Automated decision-making is appealing because artificial intelligence (AI)/machine learning (ML) systems are believed to produce more consistent, traceable, and repeatable decisions compared to humans; however, these systems come with risks that can result in discriminatory outcomes. For example, harmful biases that manifest in AI/ML-based decision systems in credit underwriting can lead to unfair results, causing impacts such as discrimination to individual applicants and potentially rippling throughout society, leading to distrust of AI-based technology and institutions that rely on it. AI/ML-based credit underwriting applications and the models and datasets that underlie them raise concerns about transparency and the identification and mitigation of bias in enterprises that seek to use ML in their credit underwriting pipeline. Yet ML models tend to exhibit "unexpectedly poor behavior when deployed in real world domains" without domain-specific constraints supplied by human operators, as discussed in NIST Special Publication (SP) 1270, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* [1]. Similar problems exist in other contexts, such as hiring and school admissions.

The heavy reliance on proxies can also be a significant source of bias in AI/ML applications. For example, in credit underwriting an AI system might be developed using input variables such as "length of time in prior employment," which might disadvantage candidates who are unable to find stable transportation, as a measurable proxy in lieu of the not directly measurable concept of "employment suitability." The algorithm might also include a predictor variable such as residence zip code, which may relate to other socio-economic factors, indirectly inferring protected attributes, and potentially resulting in the erroneous ranking of certain societal groups lower for receiving credit. This in turn would cause AI/ML systems to contribute to biased outcomes. For further information about how the use of proxies may lead to negative consequences in other contexts, see NIST SP 1270 [1].

Bias in AI systems is endemic across many application domains and can lead to harmful impacts regardless of intent. The purpose of this project is to develop domain-specific testing, evaluation, verification, and validation (TEVV) guidance for detecting bias; recommendations for managing and mitigating bias; and recommended practices for human decision makers that interact with AI-based decision systems in the specific context of consumer and small business credit underwriting. These practices can help promote fairer outcomes for those that may currently be negatively impacted by AI-based decision systems—see [1], [2]. In addition, the project aims to study the interactions between bias and cybersecurity, with the goal of identifying approaches which might mitigate risks that exist across these two critical disciplines.

This project will focus on operational, real-world AI-based decision systems, bias-detection, and bias-mitigation tools. The recommended solution architecture and practices may utilize proprietary vendor products as well as commercially viable open-source solutions. Additionally, the use and application of the NIST Dioptra test platform for testing bias will be investigated with the potential for new extensions providing new insights into the properties of an AI system. The project will include practice descriptions in the form of papers, playbook generation, and implementation demonstrations, which aim to improve the ability and efficiency of organizations to safely and securely deploy AI/ML-based decision-making technology in a specific context of interest. This project will also result in a publicly available NIST AI/ML Practice Guide, a detailed implementation guide of the practical steps needed to implement a reference design that addresses this challenge.

### Scope

The initial scope of this project is the consumer and small business credit underwriting use cases. The project will develop appropriate extensions based on third-party tools for automated bias detection and mitigation in this context of interest within the NIST Dioptra test platform. Since fairness metrics are context-specific, it is necessary to develop techniques for identifying metrics and optimizing any tradeoffs within a given real-world context (e.g., credit underwriting) and for contextually assessing gaps in current fairness metrics and processes. The project seeks approaches for how to integrate a variety of contextual factors into the ML pipeline and evaluate how humans reason and make decisions based on model output.

### Assumptions/Challenges

The following components and assumptions about them are critical for this project:

1. Dioptra, an extensible framework for AI system testing and evaluation. See the high-level architecture described in Section 3. We hope to integrate existing tools and capabilities into the Dioptra framework.
2. Third-party tools for bias detection in context. We are seeking automated tools for unwanted bias detection.
3. Test data. We are seeking appropriately defined applicant data, curated by external experts, to be used as test data.
4. AI/ML models for credit underwriting decisions along with training datasets. We are seeking third-party commercial models from willing collaborators.
5. Human subjects acting on model output as decision makers in carefully constructed trials for a specific context.
6. An end-to-end AI/ML-assisted credit underwriting decision system. We are seeking to assemble a context-specific system from components 1 through 5 and evaluate it for detecting harmful impacts stemming from unwanted bias.

### Background

NIST developed SP 1270, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* [1], as part of the AI Risk Management Framework [3], which proposes a socio-technical approach to managing risks from AI-based systems. SP 1270 provides the background for tackling harmful bias within the domain of credit underwriting.

## 2 SCENARIOS

### Scenario 1: Pre-process dataset analysis for detecting and managing bias

The goal for this scenario is transforming the data so that any underlying discrimination is appropriately treated. This method can be used if a modeling pipeline is allowed to modify the training data. This scenario will identify techniques, based on the utilization of third-party tools, and recommended practices for accomplishing mitigation. See Figure 1 and SP 1270 [1] for specific details about algorithmic approaches to bias and fairness. It is important to recognize that in the case of consumer credit underwriting, there are many legal/regulatory perspectives about whether particular approaches to debiasing are appropriate. This is usually related to legal standards on disparate treatment and disparate impact under the Equal Credit Opportunity Act [2]. This project will identify techniques and recommend practices within the legal boundaries of the law and existing regulations in the U.S.

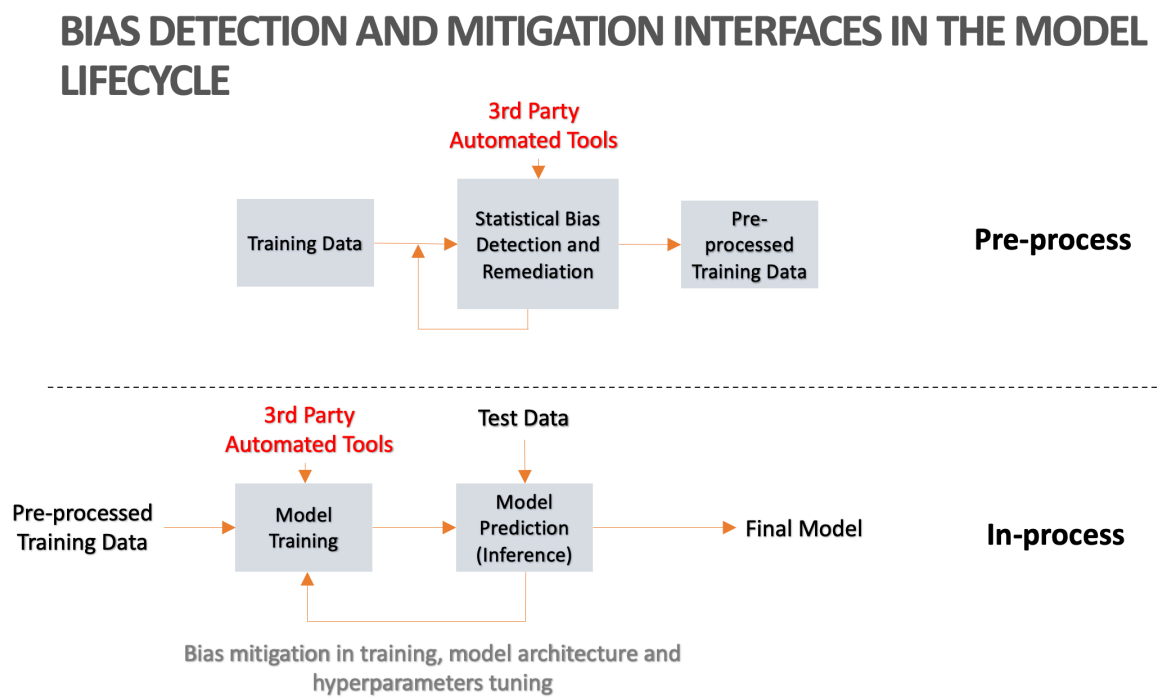


**Figure 1: Pre-Process and In-Process Workflows for Scenarios 1 and 2**

### Scenario 2: In-process model training analysis for identifying and managing statistical bias

This scenario will identify techniques, based on the utilization of third-party automated tools, and recommended practices for algorithmic modifications to manage bias during model training. Model training processes could incorporate changes to the objective (cost) function or impose a new optimization constraint. See Figure 1 and SP 1270 [1] for details. As noted in Scenario 1, there are a variety of perspectives about the appropriate way to meet legal/regulatory standards related to discrimination. This project will identify techniques and recommend practices within these legal boundaries and existing regulations in the U.S.

### Scenario 3: Post-process model inference analysis for identifying and managing statistical bias

In this scenario the learned model is treated as an opaque system and its predictions are altered by a function during the post-processing phase. The function is deduced from the performance of this opaque model on a holdout dataset (data not used in the training of the model). We will identify techniques and best practices for accomplishing this goal. See Figure 2 and SP 1270 for details.

### Scenario 4: Human-in-the-loop (HITL) decision flow for identifying and managing cognitive bias

In this scenario the output model with resulting computationally managed bias from the preceding three scenarios is presented to a human for a decision-making task specific to the credit underwriting domain. The goal here is to examine how the human interacts with AI system output to identify additional biases that may stem from this interaction and suggest strategies for effective mitigations. See Figure 2 and SP 1270 for details about human factors in AI-based decision making.

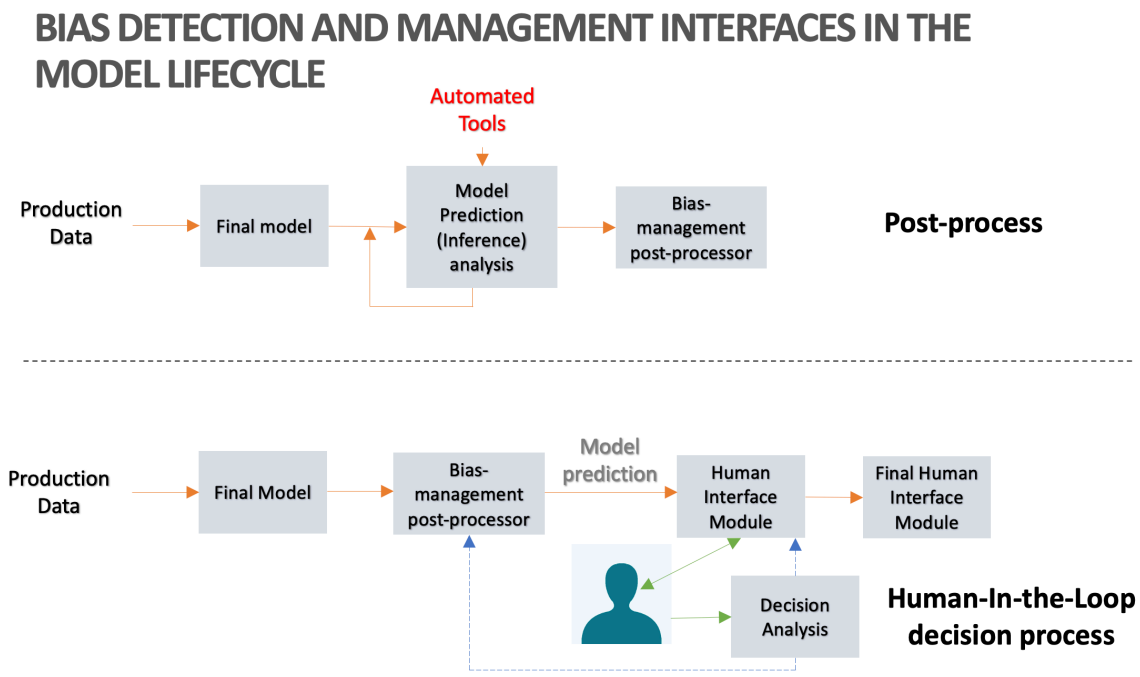


Figure 2: Post-Process and HITL Decision Process Workflows

## 3 High-Level Architecture

The high-level architecture of Dioptra is shown in Figure 3. This architecture is general and can accommodate needed extensions of the supported workflows through the MLFlow Tracking Service to support all scenarios from the previous section. The Dioptra framework will be used as the platform within which to integrate third-party bias-detection and bias-management tools and techniques.

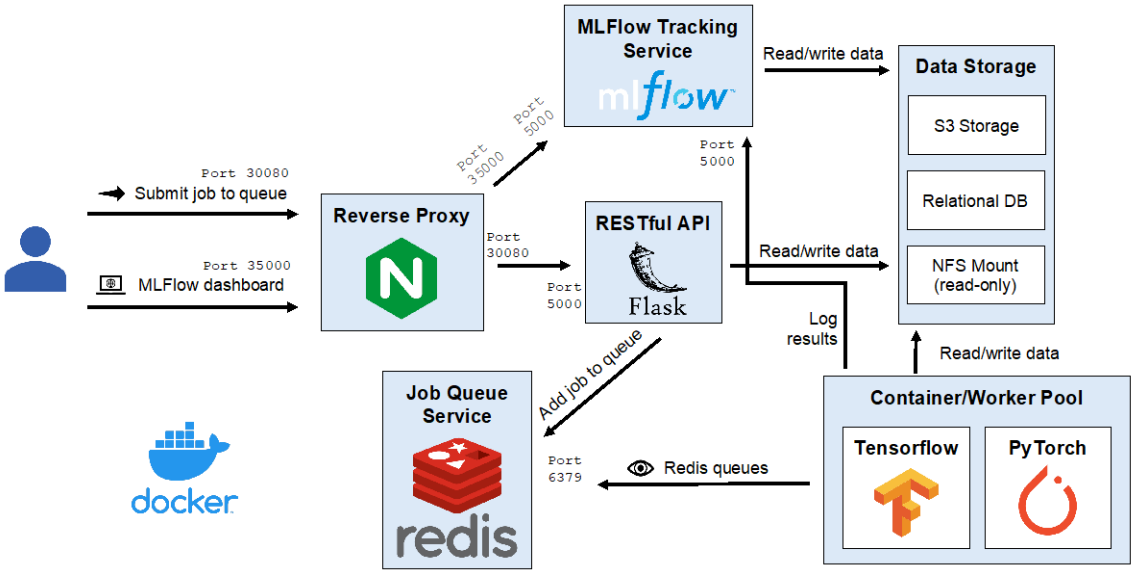


Figure 3: Dioptra High-Level Architecture

### Desired Requirements

This project aims to develop a set of bias detection and management capabilities for the typical ML workflow in the credit underwriting domain. This includes a flexible user interface (UI) to enable different simulation configurations for testing human interaction with AI decision system output and potential emergent biases in such settings.

## 4 Relevant Standards and Guidance

- NIST Special Publication 1270, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," https://doi.org/10.6028/NIST.SP.1270

## APPENDIX A REFERENCES


[1] R. Schwartz et al., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NIST Special Publication (SP) 1270, March 2022, 86 pp. Available: https://doi.org/10.6028/NIST.SP.1270.

[2] Equal Credit Opportunity Act. Available: https://www.ftc.gov/legal-library/browse/statutes/equal-credit-opportunity-act.

[3] *AI Risk Management Framework: Second Draft*, NIST, August 18, 2022, 36 pp. Available: https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

## APPENDIX B ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **HITL** | Human-in-the-Loop |
| **ML** | Machine Learning |
| **NCCoE** | National Cybersecurity Center of Excellence |
| **NIST** | National Institute of Standards and Technology |
| **SP** | Special Publication |
| **TEVV** | Testing, Evaluation, Verification, and Validation |
| **UI** | User Interface |