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

DRAFT

August 2022

ai-bias@nist.gov

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 challenges. Through this collaboration, the NCCoE develops modular, adaptable example cybersecurity 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/.

This document describes a problem that is relevant to many industry sectors. NCCoE cybersecurity experts will address this challenge through collaboration with a Community of Interest, including vendors of cybersecurity solutions. The resulting reference design will detail an approach that can be incorporated across multiple sectors.

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 comprehensive 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 guidance for recommended practices in deploying automated decision-making supported by AI/ML systems in a sector of the industry. A small but novel part of this project will be to look at the interplay between bias and cybersecurity and how they interact with each other. 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. We intend to consider other application use cases, such as hiring and school admissions, in the future. 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.

Comments on NCCoE Documents

Organizations are encouraged to review all draft publications during public comment periods and provide feedback. All publications from NIST's National Cybersecurity Center of Excellence are available at https://www.nccoe.nist.gov/.

Comments on this publication may be submitted to ai-bias@nist.gov

Public comment period: August 18, 2022 to September 16, 2022

TABLE OF CONTENTS

1 Executive Summary

Purpose

Automated decision-making is appealing because artificial intelligence (AI)/machine learning (ML) systems produce more consistent, traceable, and repeatable decisions compared to humans; however, these systems come with risks that can result in discriminatory outcomes. For example, unmitigated bias that manifests in AI/ML systems used to support automated decision making in credit underwriting can lead to unfair results, causing harms 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 decision technologies and the models and datasets that underlie them create transparency challenges generally, and those raise particular concerns about identification and mitigation of bias in enterprises that seek to use machine learning 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, and may result in ranking certain groups lower in desirability for credit approval. This in turn would cause AI/ML systems to contribute to biased outcomes. Similar issues exist in other contexts. 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 bias mitigation; and recommended practices for humans involved in automated decision-making processes in a specific context (consumer and small business credit underwriting). These practices will help promote fair and positive outcomes that benefit users of AI/ML services, the organizations that deploy them, and all of society – see [1], [3]. In addition, some attention will be given to the interactions between bias and cybersecurity, with the goal of identifying approaches which might mitigate risks that exist across these two critical characteristics of trustworthy AI.

This project will focus on operational, real-world decision automation, 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 to this area will be investigated with the potential for the addition of 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, with consideration for hiring and school admissions in future phases. The project will develop appropriate extensions based on third-party tools for automated bias detection and mitigation in a context of interest (e.g., credit underwriting, hiring, college admissions) within the NIST Dioptra test platform. Since fairness metrics are context-specific, it is necessary to identify techniques for optimizing selection of metrics within the real-world context of credit underwriting and assess gaps in current fairness metrics and processes. The project seeks approaches for how to integrate context in the ML pipeline and evaluate how humans reason and make decisions from model output in context.

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.
2. Third-party tools for bias detection in context. We are seeking automated tools for unwanted bias detection.
3. Third-party tools for bias mitigation in context. We are seeking automated tools for unwanted bias mitigation.
4. Appropriately defined applicant data, curated by external experts, used as test data.
5. AI/ML models for credit underwriting decisions along with training datasets. We are seeking third-party commercial models from willing collaborators.
6. Human subjects acting on model output as decision makers in carefully constructed trials for a specific context.
7. An end-to-end AI/ML-assisted credit underwriting decision system. We are seeking to assemble a context-specific system from components 1 through 6 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 [2], which proposes a comprehensive socio-technical approach to mitigating bias in AI and articulates the importance of context in such endeavors. This document provides the background for the project.

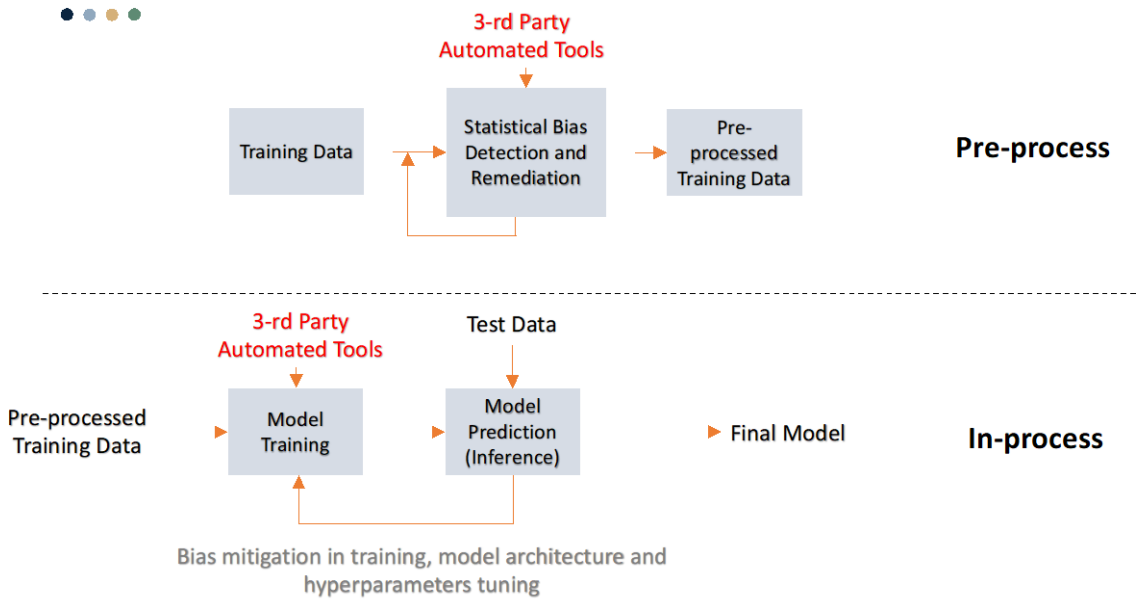
2 Scenarios

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

The goal for this scenario is transforming the data so that the underlying discrimination is mitigated to the extent possible. 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 details. It is important to recognize that in the case of consumer credit underwriting, there exists legal/regulatory ambiguity about whether particular approaches to debiasing are

appropriate, given prohibitions on disparate treatment and disparate impact under the Equal Credit Opportunity Act [3]. This project will identify techniques and recommend practices within the legal boundaries of the law and existing regulations.

Figure 1: Pre-Process and In-Process Workflows



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

This scenario will identify techniques, based on the utilization of third-party automated tools, and recommended practices that modify the algorithms in order to mitigate 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 we noted in Scenario 1, there exists some legal/regulatory ambiguity about whether particular approaches to debiasing are appropriate, given prohibitions on disparate treatment and disparate impact under the Equal Credit Opportunity Act [3]. This project will identify techniques and recommend practices within the legal boundaries of the law and existing regulations.

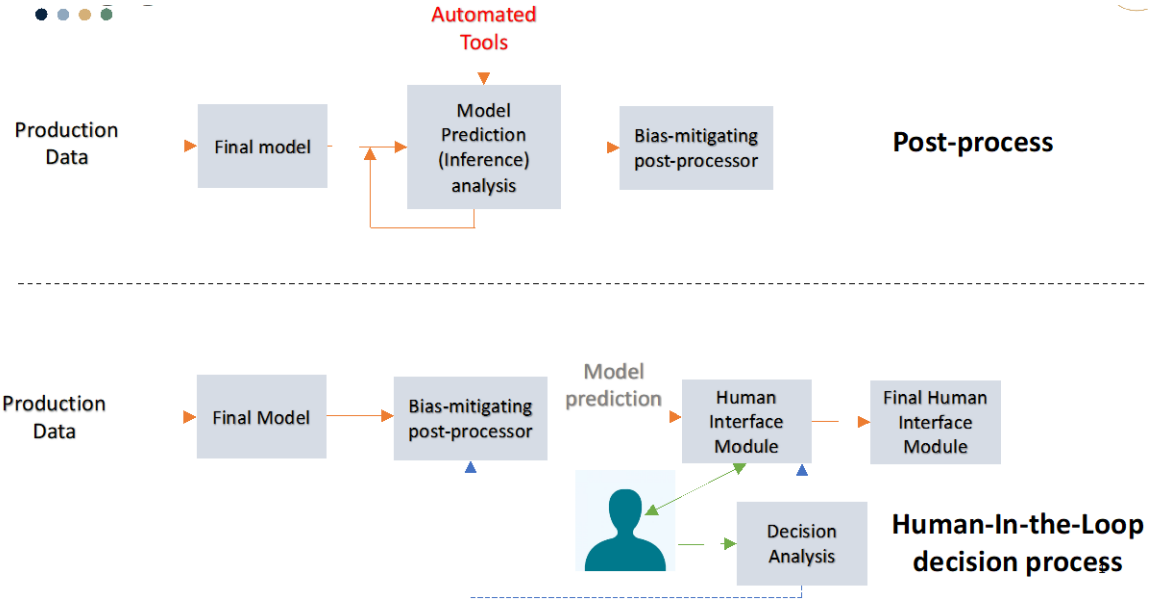
Scenario 3: Post-process model inference analysis for identifying and mitigating statistical bias

In this scenario the learned model is treated as a black box and its predictions are altered by a function during the post-processing phase. The function is deduced from the performance of the black-box model on the holdout dataset. This scenario is typically performed with the help of 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 mitigating cognitive bias

In this scenario the trained and debiased model from the preceding three scenarios is used to assist a human in making a decision specific to the context of use: credit underwriting, or other context. The goal here is to examine the different ways the human and the machine interact to detect bias stemming from this dynamic coupling of two potentially biased entities and suggest strategies for effective mitigations. See Figure 2 and SP 1270 for details.

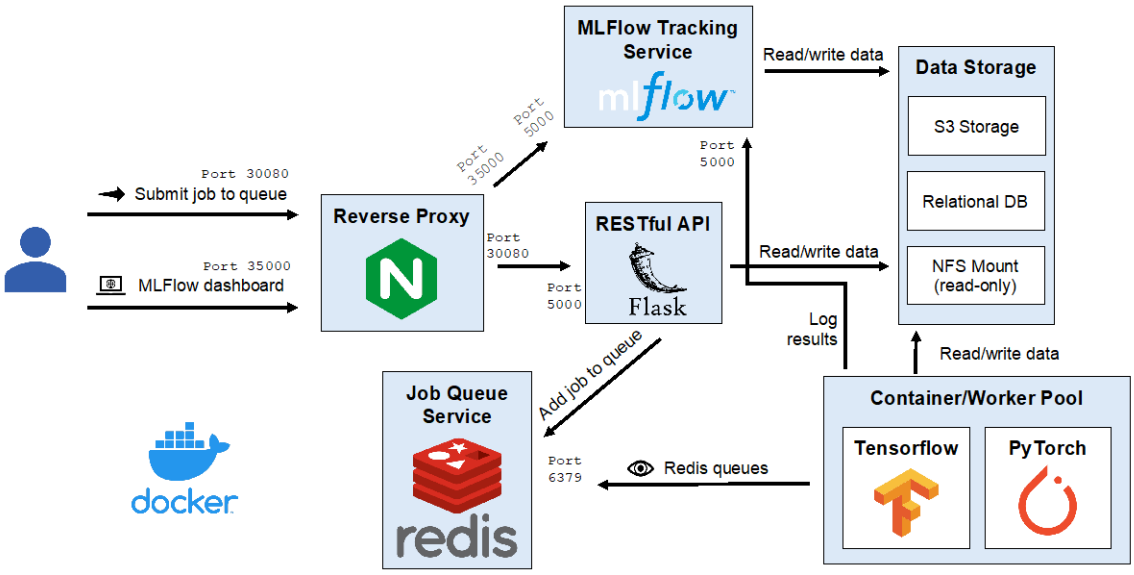
Figure 2: Post-Process and HITL Decision Process Workflows



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 ML Tracking Flow Service to support all scenarios from the previous section. The Dioptra framework will be used as the platform in which to integrate third-party bias-detection and bias-mitigation tools and techniques.

Figure 3: Dioptra High-Level Architecture



Desired Requirements

This project aims at building a set of automated bias detection and mitigation capabilities closely aligned with the typical ML workflow, including a flexible user interface (UI) component that

allows different configurations for simulating various scenarios of interaction with the HITL to enable effective detection of potential decision biases resulting from the interaction between the human and the machine.

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] *AI Risk Management Framework: Initial Draft*, NIST, March 17, 2022, 23 pp. Available: https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf

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

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