
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



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

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

12 **ABSTRACT**

13 Managing bias in an AI system is critical to establishing and maintaining trust in its operation.
14 Despite its importance, bias in AI systems remains endemic across many application domains
15 and can lead to harmful impacts regardless of intent. Bias is also context-dependent. To tackle
16 this complex problem, we adopt a comprehensive socio-technical approach to testing,
17 evaluation, verification, and validation (TEVV) of AI systems in context. This approach connects
18 the technology to societal values in order to develop guidance for recommended practices in
19 deploying automated decision-making supported by AI/ML systems in a sector of the industry. A
20 small but novel part of this project will be to look at the interplay between bias and
21 cybersecurity and how they interact with each other. The project will leverage existing
22 commercial and open-source technology in conjunction with the NIST Dioptra, an
23 experimentation test platform for ML datasets and models. The initial phase of the project will
24 focus on a proof-of-concept implementation for credit underwriting decisions in the financial
25 services sector. We intend to consider other application use cases, such as hiring and school
26 admissions, in the future. This project will result in a freely available NIST AI/ML Practice Guide.

27 **KEYWORDS**

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

31 **DISCLAIMER**

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

37 **COMMENTS ON NCCoE DOCUMENTS**

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

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

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

43 **TABLE OF CONTENTS**

44 **1 Executive Summary 3**

45 Purpose 3

46 Scope..... 4

47 Assumptions/Challenges..... 4

48 Background 4

49 **2 Scenarios 4**

50 Scenario 1: Pre-process dataset analysis for detecting and mitigating bias..... 4

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

52 bias 5

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

54 bias 5

55 Scenario 4: Human-in-the-loop (HITL) decision flow for identifying and mitigating cognitive

56 bias 5

57 **3 High-Level Architecture 6**

58 Desired Requirements 6

59 **4 Relevant Standards and Guidance 7**

60 **Appendix A References 8**

61 **Appendix B Acronyms and Abbreviations 9**

62 1 EXECUTIVE SUMMARY

63 Purpose

64 Automated decision-making is appealing because artificial intelligence (AI)/machine learning
65 (ML) systems produce more consistent, traceable, and repeatable decisions compared to
66 humans; however, these systems come with risks that can result in discriminatory outcomes.
67 For example, unmitigated bias that manifests in AI/ML systems used to support automated
68 decision making in credit underwriting can lead to unfair results, causing harms to individual
69 applicants and potentially rippling throughout society, leading to distrust of AI-based technology
70 and institutions that rely on it. AI/ML-based credit underwriting decision technologies and the
71 models and datasets that underlie them create transparency challenges generally, and those
72 raise particular concerns about identification and mitigation of bias in enterprises that seek to
73 use machine learning in their credit underwriting pipeline. Yet ML models tend to exhibit
74 “unexpectedly poor behavior when deployed in real world domains” without domain-specific
75 constraints supplied by human operators, as discussed in NIST Special Publication (SP) 1270,
76 *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* [1]. Similar
77 problems exist in other contexts, such as hiring and school admissions.

78 The heavy reliance on proxies can also be a significant source of bias in AI/ML applications. For
79 example, in credit underwriting an AI system might be developed using input variables such as
80 “length of time in prior employment,” which might disadvantage candidates who are unable to
81 find stable transportation, as a measurable proxy in lieu of the not directly measurable concept
82 of “employment suitability.” The algorithm might also include a predictor variable such as
83 residence zip code, which may relate to other socio-economic factors, and may result in ranking
84 certain groups lower in desirability for credit approval. This in turn would cause AI/ML systems
85 to contribute to biased outcomes. Similar issues exist in other contexts. For further information
86 about how the use of proxies may lead to negative consequences in other contexts, see NIST SP
87 1270 [1].

88 Bias in AI systems is endemic across many application domains and can lead to harmful impacts
89 regardless of intent. The purpose of this project is to develop domain-specific testing,
90 evaluation, verification, and validation (TEVV) guidance for detecting bias; recommendations for
91 bias mitigation; and recommended practices for humans involved in automated decision-making
92 processes in a specific context (consumer and small business credit underwriting). These
93 practices will help promote fair and positive outcomes that benefit users of AI/ML services, the
94 organizations that deploy them, and all of society – see [1], [3]. In addition, some attention will
95 be given to the interactions between bias and cybersecurity, with the goal of identifying
96 approaches which might mitigate risks that exist across these two critical characteristics of
97 trustworthy AI.

98 This project will focus on operational, real-world decision automation, bias-detection, and bias-
99 mitigation tools. The recommended solution architecture and practices may utilize proprietary
100 vendor products as well as commercially viable open-source solutions. Additionally, the use and
101 application of the NIST Dioptra [test platform](#) to this area will be investigated with the potential
102 for the addition of new extensions providing new insights into the properties of an AI system.
103 The project will include practice descriptions in the form of papers, playbook generation, and
104 implementation demonstrations, which aim to improve the ability and efficiency of
105 organizations to safely and securely deploy AI/ML-based decision-making technology in a
106 specific context of interest. This project will also result in a publicly available NIST AI/ML Practice

107 Guide, a detailed implementation guide of the practical steps needed to implement a reference
108 design that addresses this challenge.

109 **Scope**

110 The initial scope of this project is the consumer and small business credit underwriting use
111 cases, with consideration for hiring and school admissions in future phases. The project will
112 develop appropriate extensions based on third-party tools for automated bias detection and
113 mitigation in a context of interest (e.g., credit underwriting, hiring, college admissions) within
114 the NIST Dioptra test platform. Since fairness metrics are context-specific, it is necessary to
115 identify techniques for optimizing selection of metrics within the real-world context of credit
116 underwriting and assess gaps in current fairness metrics and processes. The project seeks
117 approaches for how to integrate context in the ML pipeline and evaluate how humans reason
118 and make decisions from model output in context.

119 **Assumptions/Challenges**

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

- 121 1. [Dioptra](#), an extensible framework for AI system testing and evaluation. See the high-
122 level architecture described in [Section 3](#).
- 123 2. Third-party tools for bias detection in context. We are seeking automated tools for
124 unwanted bias detection.
- 125 3. Third-party tools for bias mitigation in context. We are seeking automated tools for
126 unwanted bias mitigation.
- 127 4. Appropriately defined applicant data, curated by external experts, used as test data.
- 128 5. AI/ML models for credit underwriting decisions along with training datasets. We are
129 seeking third-party commercial models from willing collaborators.
- 130 6. Human subjects acting on model output as decision makers in carefully constructed
131 trials for a specific context.
- 132 7. An end-to-end AI/ML-assisted credit underwriting decision system. We are seeking to
133 assemble a context-specific system from components 1 through 6 and evaluate it for
134 detecting harmful impacts stemming from unwanted bias.

135 **Background**

136 NIST developed SP 1270, *Towards a Standard for Identifying and Managing Bias in Artificial*
137 *Intelligence* [\[1\]](#), as part of the AI Risk Management Framework [\[2\]](#), which proposes a
138 comprehensive socio-technical approach to mitigating bias in AI and articulates the importance
139 of context in such endeavors. This document provides the background for the project.

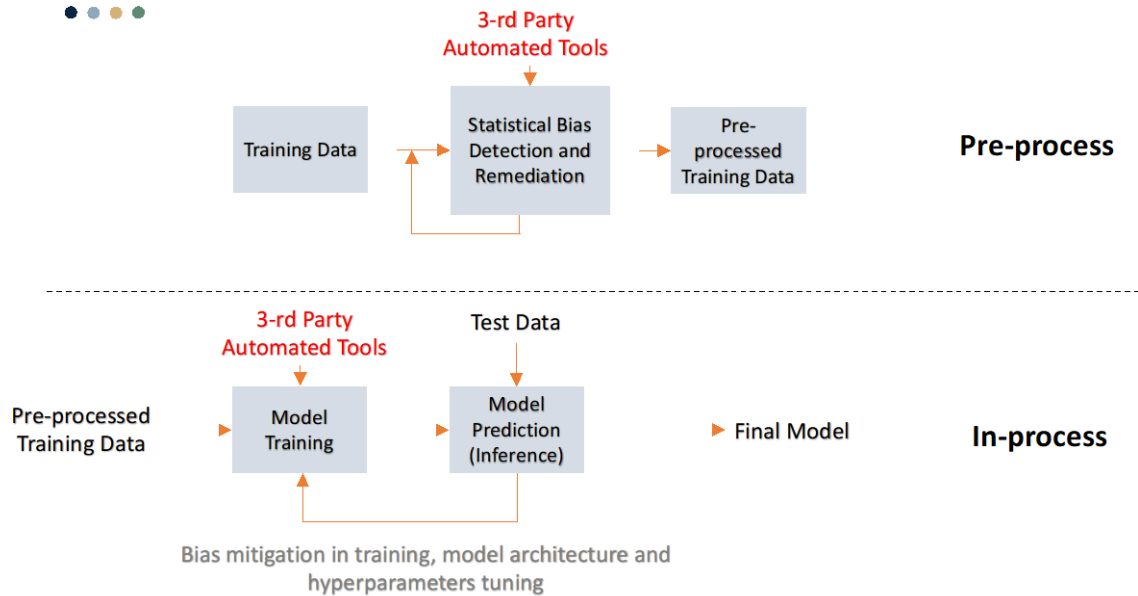
140 **2 SCENARIOS**

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

142 The goal for this scenario is transforming the data so that the underlying discrimination is
143 mitigated to the extent possible. This method can be used if a modeling pipeline is allowed to
144 modify the training data. This scenario will identify techniques, based on the utilization of third-
145 party tools, and recommended practices for accomplishing mitigation. See Figure 1 and [SP 1270](#)
146 [\[1\]](#) for details. It is important to recognize that in the case of consumer credit underwriting,
147 there exists legal/regulatory ambiguity about whether particular approaches to debiasing are

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

151 **Figure 1: Pre-Process and In-Process Workflows**



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

153 This scenario will identify techniques, based on the utilization of third-party automated tools,
 154 and recommended practices that modify the algorithms in order to mitigate bias during model
 155 training. Model training processes could incorporate changes to the objective (cost) function or
 156 impose a new optimization constraint. See Figure 1 and [SP 1270 \[1\]](#) for details. As we noted in
 157 Scenario 1, there exists some legal/regulatory ambiguity about whether particular approaches
 158 to debiasing are appropriate, given prohibitions on disparate treatment and disparate impact
 159 under the Equal Credit Opportunity Act [3]. This project will identify techniques and recommend
 160 practices within the legal boundaries of the law and existing regulations.

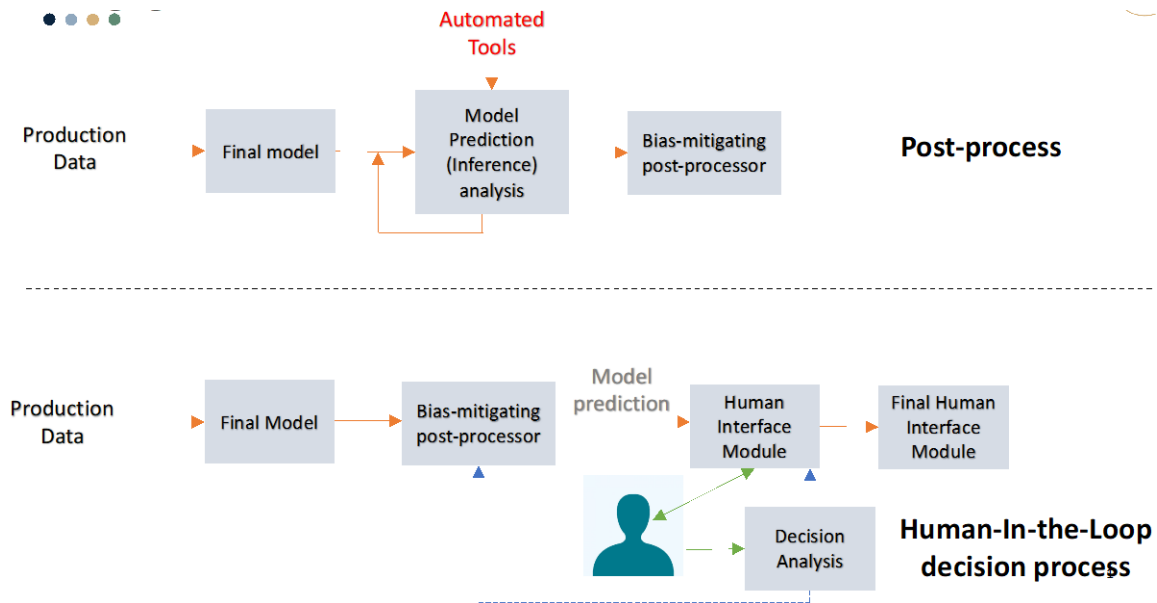
161 **Scenario 3: Post-process model inference analysis for identifying and mitigating statistical bias**

162 In this scenario the learned model is treated as a black box and its predictions are altered by a
 163 function during the post-processing phase. The function is deduced from the performance of the
 164 black-box model on the holdout dataset. This scenario is typically performed with the help of a
 165 holdout dataset (data not used in the training of the model). We will identify techniques and
 166 best practices for accomplishing this goal. See Figure 2 and [SP 1270](#) for details.

167 **Scenario 4: Human-in-the-loop (HITL) decision flow for identifying and mitigating cognitive
 168 bias**

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

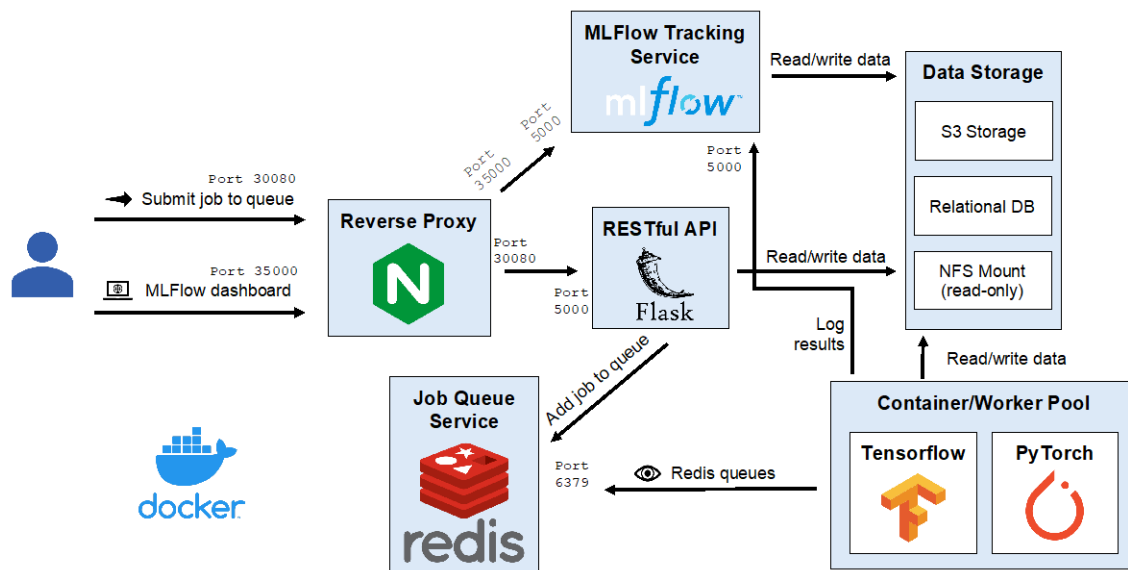
174 **Figure 2: Post-Process and HITL Decision Process Workflows**



175 **3 HIGH-LEVEL ARCHITECTURE**

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

181 **Figure 3: Dioptra High-Level Architecture**



182 **Desired Requirements**

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

DRAFT

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

188 **4 RELEVANT STANDARDS AND GUIDANCE**

- 189 • NIST Special Publication 1270, “Towards a Standard for Identifying and Managing Bias in
190 Artificial Intelligence,” <https://doi.org/10.6028/NIST.SP.1270>

191 **APPENDIX A REFERENCES**

- 192 [1] R. Schwartz et al., *Towards a Standard for Identifying and Managing Bias in Artificial*
193 *Intelligence*, NIST Special Publication (SP) 1270, March 2022, 86 pp. Available:
194 <https://doi.org/10.6028/NIST.SP.1270>
- 195 [2] *AI Risk Management Framework: Initial Draft*, NIST, March 17, 2022, 23 pp. Available:
196 <https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf>
- 197 [3] Equal Credit Opportunity Act. Available: [https://www.ftc.gov/legal-](https://www.ftc.gov/legal-library/browse/statutes/equal-credit-opportunity-act)
198 [library/browse/statutes/equal-credit-opportunity-act](https://www.ftc.gov/legal-library/browse/statutes/equal-credit-opportunity-act)

199

200 **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